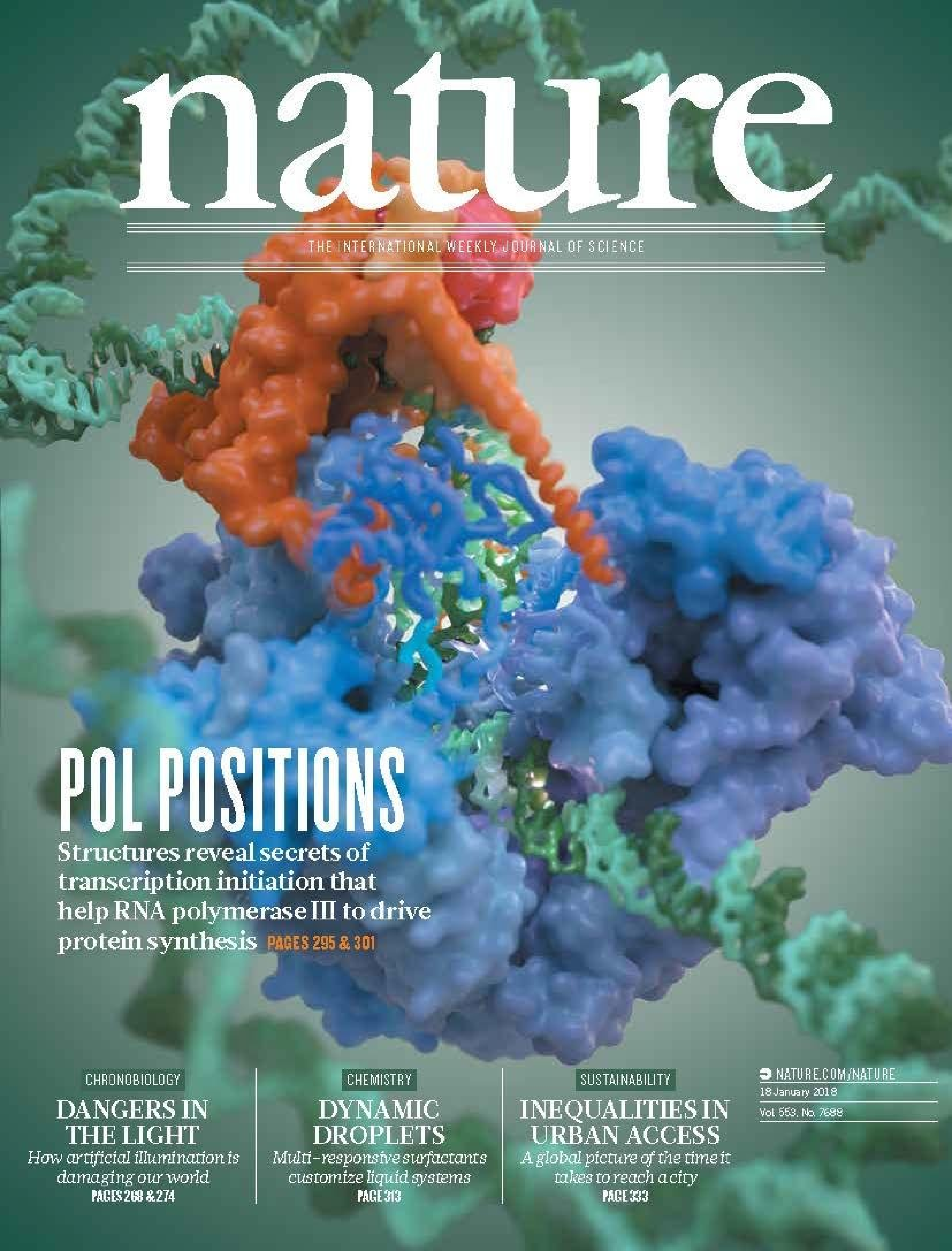
nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



POL POSITIONS

Structures reveal secrets of transcription initiation that help RNA polymerase III to drive protein synthesis PAGES 295 & 301

CHRONOBIOLOGY

DANGERS IN THE LIGHT

How artificial illumination is damaging our world

PAGES 268 & 274

CHEMISTRY

DYNAMIC DROPLETS

Multi-responsive surfactants customize liquid systems

PAGE 313

SUSTAINABILITY

INEQUALITIES IN URBAN ACCESS

A global picture of the time it takes to reach a city

PAGE 333

➔ NATURE.COM/NATURE

18 January 2018

Vol. 553, No. 7688

THIS WEEK

EDITORIALS

VOTING Maths researchers help overturn partisan electoral fiddling **p.250**

WORLD VIEW Reach out for voices from developing nations **p.251**

SANCTUARY Peru sets up nature reserve in Amazon rainforest **p.255**

Chinese science is ready to step up

The country seems to be on course to sail into scientific dominance, but it must listen to what researchers at home and abroad really need.

In the past 12 months or so, China has opened its first facility for research into the world's most dangerous pathogens, unveiled another world-leading telescope and turned on its first world-class neutron source. Researchers in the country have also established a neuroimaging factory to automate the highly detailed imaging of human brains.

Money has poured in, too. Chinese artificial-intelligence (AI) companies, in a crowded field, impressed international investors. Companies specializing in computer-vision technology pulled in more than US$1 billion in investment last year. Legend Biotech in Nanjing reported positive results from a clinical trial of a CAR-T therapy — showing its clout in a highly competitive field in which researchers engineer a patient's own cells and reintroduce them to treat cancer. In response, Janssen Biotech of Horsham, Philadelphia, put $350 million into further development of the therapy.

Look at most scientific indicators — publications, patents, number of researchers — and China seems to be on course to sail into scientific dominance. And, as many observers point out, that could happen much sooner than anyone previously expected if the US government continues to hold policies as destructive to science as those pushed by the administration of President Donald Trump. The upshot of this is a lot of opportunities for researchers in China. A Career Guide starting on page S1 this week offers details on how to embrace them.

But pitfalls lie in wait if officials and researchers in China are not careful. The country's AI research, for example, is booming right now, with publications outpacing those produced in the United States. But researchers acknowledge that many of these papers are not of particularly high quality. They also wonder whether Chinese academia or industry will invest in the ways necessary to create fundamental breakthroughs in the field.

As we discuss in a News story on page 260, billions of dollars announced for a provincial AI park in China came as a surprise to many AI researchers in the capital, Beijing. This doesn't bode well, because it suggests a top-down effort made without consulting the research or academic community. Existing pricey science parks dedicated to trendy fields such as biotechnology and software development have produced mixed results and raised the question of whether resources are being wasted on fancy infrastructure.

Meanwhile, China might ratchet up its firm grip on the Internet. If it does so, many scientists there could lose access to the virtual private networks that they use to bypass restrictions and reach crucial websites such as Google Scholar. That would cut off access to literature, results and discussion, and isolate them from the international community.

Despite China's claim to the throne of scientific superpower, the government retains a soft spot for unproven claims of traditional Chinese medicine. (This is one area in which the United States, in its attempts to rein in naturopathy and homeopathy in the past two years, seems to be cleaning up its own scientific house.)

The lack of transparent or predictable funding decisions could also derail China's ambitions. Although the National Natural Science Foundation of China is generally well regarded for the grants it distributes, however small, larger projects continue to be marked by disarray. Neuroscientists have been sounded out to join a multimillion-dollar national programme meant to rival (and hopefully complement) brain-science projects in the United States, Europe and Japan. But so far, all the Chinese project has produced is false starts and confusion as scientists attempt to ready their research programmes to align with a national project that is always just around the corner.

> ***"Lack of transparent funding decisions could derail China's ambitions."***

China is right to praise itself for its accomplishments in building a successful scientific community. And its stated goals of becoming an attractive place for foreign or returning scientists and a more desirable partner for international collaborations are the right ones for a country ready to take up a much needed leadership role and act as a model for other nations. But China will need to make more effort to listen to its scientists and survey the needs of researchers elsewhere to find out what problems — including those mentioned above — might hamper attainment of those goals. ■

Vaccine boosters

A new French law that makes immunizations mandatory is not the only way to improve.

It is one thing to be certain (as *Nature* is) that widespread immunization is a vital tool for public health. But it is much more contentious, given the diversity of humanity's ethical and cultural norms, to impose vaccinations on a population. That diversity is reflected, for example, by differing choices among countries in Europe: some (mostly the post-Soviet Union states) make vaccinations for many diseases mandatory, whereas the majority do not.

France is now providing a case study of exactly these debates.

A new French law requires that babies born after 1 January be vaccinated in their early years against 11 diseases. Previously, vaccines against only three of these — diptheria, tetanus and polio — were mandatory. The others were recommended, but the decision was left to parents. Now, children must also be vaccinated against *Haemophilus influenzae* B, hepatitis B, pertussis, pneumococcal disease, meningitis C, measles, mumps and rubella. Those who haven't had all their immunizations, including booster shots, the government says, will be refused admission to nurseries, schools and camps in France.

This policy is dividing public-health scientists in the country. Many French general practitioners are among those who argue that the measure is authoritarian and could backfire, not least by alienating parents and increasing wariness of vaccines in a country where various health scandals (most infamously, HIV-infected blood transfusions given in the early 1980s to people with haemophilia) have spread mistrust of health authorities.

Misguidedly, authorities seem to think that the new law is a pertinent response to scare stories about the safety of childhood vaccines, in particular, those told by anti-vaccine groups. Countering such misinformation is important, but does not alone constitute the basis for a coherent vaccine policy. Data on vaccine coverage of most diseases in France show that the situation is now better than it has been in years. Coverage rates for some newer vaccines are too low, but have nonetheless been increasing; the rates of meningitis-C vaccination, for example, have steadily increased since it was introduced a decade ago, from just 48% among 2 year olds at the end of 2011 to 71% in 2016. But vaccine coverage in France for most diseases is high overall. The challenge is rather to develop policies that will get the stragglers vaccinated to ensure that enough of the population is immunized to surpass the thresholds needed for herd immunity.

To portray societal hesitation about vaccination as a simple battle between anti-vaccine groups and ignorant populations on the one side, and scientific reason and public health on the other — as the French government has done — promotes an unproductive and sterile controversy, and a simplified view that obscures complex issues, such as the multiple causes of 'vaccine hesitancy' in populations, and the fundamental role of building trust in health-care institutions and information from government and scientists.

One of the biggest practical problems that France faces is the often poor follow-through of booster shots. Health data show that only eight in ten babies get the MMR booster (for mumps, measles and rubella) due at 18 months of age — a lower rate than in many other countries, and a problem because it weakens herd immunity in the population.

This has no-doubt contributed to a slight recrudescence of measles in the country, with a few dozen to a few hundred cases annually — and in particular, to an epidemic of several thousand cases in 2010 and 2011. But the French government's reaction of making childhood vaccines mandatory is simplistic, and reneges on the administration's greater responsibility to work patiently hand in hand with health-care workers and the public to improve what is already high take-up of vaccines. Multiple studies show that simple reminders — text messages among them — of when vaccines and booster shots are due can have a big impact on compliance and coverage. The same is true of national electronic vaccine-information systems to track people's vaccinations, an area in which much progress remains to be made.

"The challenge is to develop policies that will get the stragglers vaccinated."

To its credit, the French government has pledged to review annually the compliance and impact of the new law. But in a country where '*liberté*' is one of the three pillars of the national motto, the heavy-handed law could do something that nobody involved wants: fuel further unfounded resistance to life-saving vaccines. Making vaccines mandatory should be at most a stopgap. The only sustainable policy is for the government to put its efforts into making a strong case to the public about the benefits of vaccinations, and to better use the available evidence to implement more proactive strategies that can extend already respectable coverage rates for most diseases to those vaccines that are lagging. ■

Electoral plot

Maths helps to catch Republican politicians who unfairly fiddled with voting districts.

Mathematicians are no longer devices for turning coffee into theorems, as the Hungarian mathematics researcher (and caffeine addict) Alfréd Rényi is said to have claimed. They seem pretty useful for preserving democracy, too. In striking down the way that officials in North Carolina unfairly partitioned the state into electoral districts, a US federal court last week conspicuously cited the work of mathematicians including Jonathan Mattingly, an expert in mathematical modelling.

In a 200-page decision released on 9 January, the three-judge court in Richmond, Virginia, said that the districting had unfairly favoured the Republican Party. Maths played a key part in helping the court to reach that decision, by demonstrating the unlawful use of partisan gerrymandering — fiddling with district boundaries to include or exclude certain voters and steer the results of an election. Those apportioning districts might draw borders that pack large numbers of voters for an opposition party into a small number of districts, for example, limiting the number of seats that the opposition can win. The process has been likened to allowing lawmakers to choose their voters, rather than the other way around.

Mattingly, a researcher at Duke University in Durham, North Carolina, used his expertise to argue that the state districts were drawn up to give Republicans an unfair advantage. To do so, he used an algorithm that produced around 24,000 maps of marginally different district configurations that were randomly drawn on the basis of geographic criteria. The Republican-drawn boundaries, which had delivered 9 Republicans to the state's 13 seats in the House of Representatives in Washington DC in 2012, were more gerrymandered than practically every single one of Mattingly's algorithm-derived maps. Using the same voting data, his maps nearly all gave a larger number of wins to the Democratic Party and, in many cases, gave it the majority.

Mattingly had taken an interest in the process after the 2012 elections and was called to testify after two advocacy organizations sued the state in federal court following the 2016 elections. In October, they asked Mattingly to take the stand and explain his work and its implications. He was ready: by then, he and his collaborators had done more-recent studies of the state's current redistricting, engineered in 2016 by the Republican majority in the North Carolina General Assembly.

Some of the modelling is preliminary, but it has had a historic impact: last week's ruling was the first time that a US federal court has struck down electoral districting for favouring one political party over another. (Previous rulings have done so for other reasons, such as racial disparities.) Gerrymandering is not exclusive to North Carolina, or to the US Republican Party. Courts have struck down pro-Democratic redistricting in Maryland, for example, and similar cases are being debated in the United Kingdom and elsewhere.

Last week's ruling is not the final word on North Carolina's system. The General Assembly has filed an appeal, and the case is likely to end up in the US Supreme Court. The court has ruled in the past that politically motivated gerrymandering was illegal, but also that there were no objective metrics to establish it.

But that is what Mattingly and others have been working to change — and the computer simulations could be needed more than ever. The upcoming 2020 US census will trigger widespread redrawing of electoral districts, and there are already concerns that gerrymandering will be rife.

Mattingly and other academics who study electoral systems are organizing to train their colleagues on the science of gerrymandering, and how to communicate it to a non-mathematical audience. One summer camp held last year had planned for 50 attendees; more than 1,000 applied. That's a lot of coffee — and all of it consumed in a good cause. ■

WORLD VIEW *A personal take on events*

Showcase scientists from the global south

The contributions of researchers in the developing world must be sought and recognized, says **Dyna Rochmyaningsih**.

A new species of orangutan, mud volcanoes that bury villages, Zika virus: there is no shortage of science stories emerging from the 'global south', a group of countries across Africa, South America and Asia that endured colonialism and are now struggling to improve their economies. But we need to pay attention to which scientists are telling us this region's stories.

Researchers in the global south take part in cutting-edge research, yet their names usually fall under the shadows of scientists from the West. Although those coming to the region for research are often more extensively trained than local scientists, that is not the only reason they usually receive more credit.

At least one randomized, blinded study points to bias against researchers from the global south. In an experiment framed as a speed-reading exercise, 347 English clinicians rated the same 4 abstracts twice; each time, the abstracts were given different author affiliations. Abstracts supposedly originating from leading US and German universities scored higher than identical ones attributed to top universities in Ethiopia and Malawi (M. Harris *et al. Health Aff.* **36,** 1997–2004; 2017). The country of origin mattered more for rankings than the title of the journal. The study authors predict that research from low-income countries is "discounted prematurely and unfairly".

SCIENTISTS FROM **THE SOUTH** MIGHT BE THE 6TH OR 16TH **AUTHOR,** YET THE WORK COULD **NOT PROCEED** WITHOUT THEM.

Authorship position is another big issue. According to Danang Birowosuto, an Indonesian physicist now at Nanyang Technological University in Singapore, who has worked in scientific institutions across the globe, researchers with Indonesian affiliations are seldom listed as the first or lead author — in part because they rarely contribute the biggest slice of funding. This imbalance keeps these scientists from proposing research and developing ideas. And because the most prominent researchers have the most success in future funding cycles, the situation is self-perpetuating. It will continue unless the scientific community confronts it head on.

Better acknowledgement would help. For international projects, scientists from the south might be the 6th or 16th author, yet the work could not proceed without them.

Take the discovery of new primate species reported late last year (A. Nater *et al. Curr. Biol.* **27,** 3487–3498; 2017). It was an Indonesian scientist who collected the 500 orangutan skulls from 21 institutions around the world and did morphological analyses. It was also Indonesian scientists who mediated the complicated process of gaining research access to pristine forests. This was rarely acknowledged in news stories. Such details are fascinating, important and too often missing from the literature and related news coverage.

No wonder some relationships become strained. Last year, *The Jakarta Post* ran stories about biopiracy of specimens of insects and marine life. I know of at least one researcher who has worked with Western scientists expecting to be an author, only to find that manuscripts have been prepared, submitted and published without them being informed of any of these steps, or being listed as an author.

Because of these concerns, the Indonesian Ministry of Research, Technology and Higher Education has placed conditions on foreign scientists who hope to work in the country's pristine areas, such as in parts of the Banda Sea: an Indonesian scientist must lead the research. The goal is to make sure that expertise on science originating from Indonesia is retained in Indonesian researchers.

An example of this process at its best can be found in work on *Homo floresiensis*, a hominid discovered in 2003 and nicknamed the hobbit. The late archaeologist Michael Morwood, of the University of Wollongong in Australia, respectfully cultivated long-term collaboration with Indonesian researchers and, because of this, succeeded in gaining access to sites. The first papers on the hominid rightfully gave Indonesian authors prominent positions; the researchers have thrived since.

Collaborators should set clear expectations and encourage local researchers to participate in tasks that will be formally recognized. Western scientists should also explicitly solicit input: southern scientists can be reluctant to critique study designs for fear of disturbing collaborations. Southern scientists must look beyond local networks and proactively seek international funding.

When publishing science from the south, senior Western authors can take simple steps to share credit. They should provide short descriptions of authors' contributions. When they give interviews, they should emphasize the roles performed by local scientists, and encourage journalists to interview them.

Journalism should also strive to include more scientists from the south, both those who participated in the research and those who are in a unique position to provide comment. In one story I reported on, about a tropical disease caused by roundworms, an interview with an Indonesian scientist showed how World Health Organization recommendations would have limited effectiveness because they were not tailored to species differences across the islands.

SciDev.Net, a website that covers science and technology in the developing world, led the way in fostering such skills and connections. 'The Conversation,' a website established by a group of universities, provides a good platform for southern scientists to air their views.

We must all work together to bring scientists from the global south out of the shadows. ■

Dyna Rochmyaningsih *is a science journalist in Medan, Indonesia.*
e-mail: drochmya87@gmail.com

SEVEN DAYS *The news in brief*

EVENTS

Meeting probe

University College London (UCL) has launched an internal inquiry after it emerged that a series of controversial conferences on intelligence took place there. The meeting, called the London Conference on Intelligence, has been held annually since at least 2015 and was organized by psychologist and UCL honorary lecturer James Thompson, the university said. Talk topics have included purported disparities in cognitive ability linked to genetics, race and gender, and the cover of the 2016 conference brochure features an early-twentieth-century quote by a US eugenics proponent. UCL said in a statement that it had not approved the events and its officials had not been informed about the meetings' speakers, as they should have been. It is investigating a possible breach of its room-booking process, and said that it is "committed to free speech but also to combatting racism and sexism in all forms". Thompson did not respond to requests for comment. The details of the meetings were reported last week by the *London Student* newspaper and the UK magazine *Private Eye*.

XINHUA VIA ZUMA

Tanker crash raises oil-spill fears

An Iranian oil tanker sank in the East China Sea on 14 January, eight days after colliding with a cargo ship. The *Sanchi* was carrying 136,000 tonnes of ultra-light crude oil, and had been adrift and partly on fire after the accident, prompting fears of a significant oil spill. Chinese state media reported that the fire spread further and the tanker sank. The vessel's 32 crew perished in the accident. It is so far unclear how much oil has been released. China's State Oceanic Administration says that it will monitor the site for environmental damage.

FACILITIES

Maths institute

Imperial College London and the French CNRS — Europe's largest basic-science agency — inaugurated a joint mathematics laboratory at the British university's campus on 15 January. The centre, named after French mathematician Abraham de Moivre, is the first CNRS research unit to be established in the United Kingdom, and will give Imperial College mathematicians continued access to French funding after the United Kingdom leaves the European Union in 2019. It builds on a fellowship programme that supports extended stays by French mathematicians at Imperial, and will also sponsor Imperial mathematicians to spend time at French institutions.

PEOPLE

Society head

Former US Department of Energy secretary Steven Chu is the president-elect of the American Association for the Advancement of Science (AAAS), the organization announced on 9 January. He succeeds former US Food and Drug Administration commissioner Margaret Hamburg, who will begin her term as the group's president in February. AAAS leaders spend a year as president-elect, a year as president and a year as chair of the board of directors. Chu is currently a physicist at Stanford University in California and is known for his work developing ways to cool and trap atoms using lasers, which won the 1997 Nobel Prize in Physics.

Indian space chief

Renowned space scientist K. Sivan has been appointed chief of the Indian Space Research Organisation (ISRO). Sivan, who is currently the director of ISRO's Vikram Sarabhai Space Centre in Thiruvananthapuram, replaces A. S. Kiran Kumar, whose three-year tenure ended on 14 January. Sivan joined ISRO in 1982 and has contributed to the design and mission planning of two of the agency's key launch systems: the Geosynchronous Satellite Launch Vehicle and the Polar Satellite Launch Vehicle. On 12 January, the agency launched its 100th satellite, along with 30 others.

POLICY

Fossil-fuel funds

New York City's pension funds will shed investments worth some US$5 billion distributed between more than 190 fossil-fuel companies over the next 5 years. Mayor Bill de Blasio and other officials announced the decision on 10 January. De Blasio also announced that the city has filed a lawsuit against five of the largest publicly traded fossil-fuel companies: BP, Chevron, ConocoPhillips, Exxon Mobil and Royal Dutch Shell. The suit seeks billions of dollars in damages for harm that the city has already sustained as a result of global warming, as well as future spending to address the effects of climate change.

FUNDING

Italian excellence

The Italian ministry of research announced on 9 January how it will distribute a €1.3-billion (US$1.6-billion) pot of funds as part of an initiative aimed at boosting research excellence. The money will be given out over 5 years to 180 university departments judged in a competition to have the strongest research plans. It is the first time Italy has run a competitive research-excellence scheme. Just 25 winners were from universities in the poorer south of the country. The ministry pledged to use €110 million of European Union subsidies to bolster prospects in future excellence competitions for researchers in disadvantaged areas.

POLITICS

Immigration lawsuit

The US government must continue a programme that gives temporary residency to people who entered the United States illegally as children, a federal district court judge in San Francisco, California, said on 9 January. US President Donald Trump had sought to end the programme, called Deferred Action for Childhood Arrivals (DACA), with his administration announcing last September that it would not renew work permits awarded through DACA after 5 March, leaving 800,000 young immigrants, including some scientists, in limbo. In response, several states sued the federal government; the lawsuit that prompted the new injunction was filed by the state of California and the University of California system. The nationwide injunction will stay in place while the case wends its way through the legal system.

Map mayhem

Mathematical analyses featured heavily in a 9 January US federal-court decision that found a North Carolina congressional-district map unconstitutional. The map defines voting districts that will each elect one member to send to the House of Representatives this year. A panel of three judges ruled unanimously that it had been drawn to give an advantage to the Republican Party. Among other lines of evidence, the judges cited work by mathematician Jonathan Mattingly of Duke University in Durham, North Carolina, in their conclusion that lawmakers had crafted the map "to subordinate the interests of non-Republican voters". The state has until 29 January to submit a redrawn map to the court. See page 250 for more.

RESEARCH

Dark-energy data

The Dark Energy Survey — an effort to probe the properties of the mysterious force that is accelerating the expansion of the Universe — made its first three years of data freely available on 10 January. It is the first major release by the project, which launched in 2013. The data contain information on about 400 million astronomical objects. Cosmologists have already used some of the information to create the biggest map yet of the Universe, charting the distribution of matter in part by measuring how mass bends light. The survey is a collaboration of more than 400 researchers, and mainly gathers data using a telescope in Chile.

CONSERVATION

Rainforest park

JONH JAIRO MUESES-CISNEROS

Peru has designated a new reserve in the heart of the Amazon rainforest. The Yaguas National Park, announced on 11 January, covers nearly 869,000 hectares along the Putumayo River in northeastern Peru. The area houses two-thirds of the country's freshwater fish species as well as thousands of plants, birds and other animals (**pictured**, a *Hypsiboas* frog), according to scientists at the Field Museum in Chicago, Illinois, who led a team that surveyed the area in partnership with local communities and the Peruvian government. As well as protecting the forest and wildlife, scientists say, the park will benefit indigenous residents by helping to prevent illegal logging and gold mining that harm their health and livelihoods.

TECHNOLOGY

AI beats humans

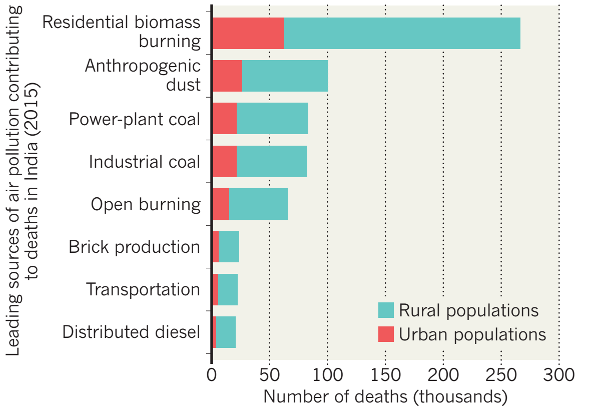
Artificial-intelligence programs built by Chinese e-commerce giant Alibaba and Microsoft have scored higher than humans in a challenging reading-comprehension test. On 15 January, Alibaba announced that the company's deep-neural-network model had scored 82.44 on the Stanford Question Answering Dataset, a machine reading-comprehension test in which machines must provide exact answers to questions based on 500 Wikipedia articles. Microsoft's model scored 82.65, also beating human performance (82.304). Alibaba's chief scientist of natural-language processing, Luo Si, said that the company's AI technology could be used for customer service, museum tutorials and medical enquiries.

TREND WATCH

Air pollution was responsible for about 1.1 million deaths in India in 2015, according to a report by an Indian collaboration released on 11 January. Burning fuel at home contributed to the deaths of 268,000 people — roughly 25% of all deaths caused by inhalation of fine airborne particles. India has some of the world's worst air pollution, with almost all of its population living in areas where pollution exceeds World Health Organization guidelines.

DEATHS FROM AIR POLLUTION IN INDIA

In 2015, the burning of biomass in homes contributed to the deaths of nearly 270,000 people in India, most of them living in rural areas.



SOURCE: HEALTH EFFECTS INSTITUTE

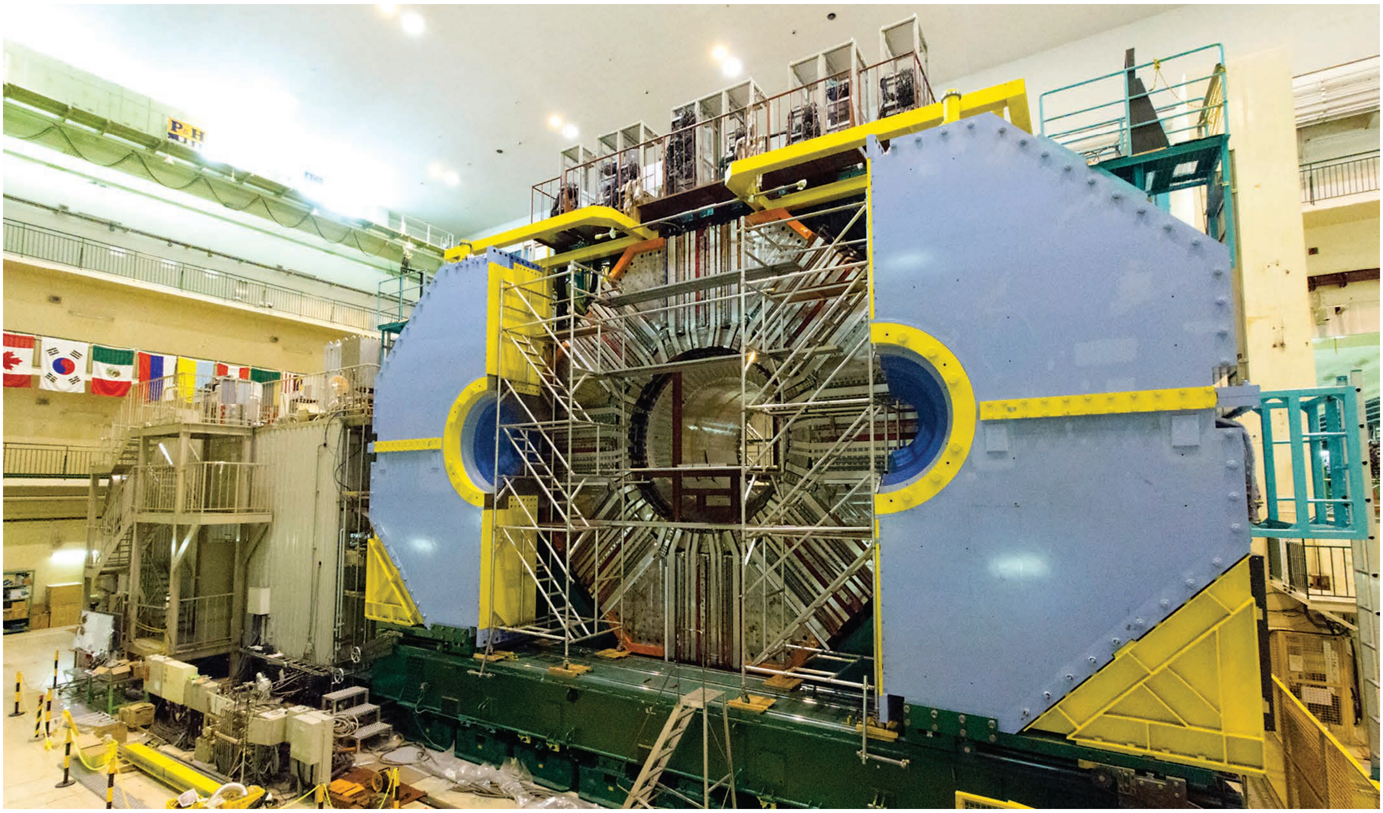
➲ NATURE.COM
For daily news updates see:
www.nature.com/news

GENETICS Edited genes could keep modified organisms from spreading **p.259**

COMPUTING Labour shortage may limit China's push in artificial intelligence **p.260**

SPACE Experiment tests pulsars as interstellar navigation aid **p.261**

ECOLOGY The new fight over the light that steals the night **p.268**



KEK/BELLE II

The Belle II experiment at the High Energy Accelerator Research Organization (KEK) in Tsukuba, Japan.

PARTICLE PHYSICS

Collider seeks cracks in physics framework

The Belle II experiment in Japan will search for missing pieces in the standard model.

BY ELIZABETH GIBNEY

The quest to explore the frontiers of physics will heat up in Japan next month, when beams of high-energy electrons are set to start smashing into their antimatter counterparts at one of the world's premier accelerator laboratories. The experiment, called Belle II, aims to chase down rare, promising hints of new phenomena that would extend the standard model — a remarkably successful, but incomplete, physics theory that describes matter and forces.

In February, an accelerator at Japan's High Energy Accelerator Research Organization (KEK) in Tsukuba will begin an initial six-month run of collisions. The eventual goal is to chart in high precision the decays of B-mesons, which contain a fundamental building block of nature known as a the b quark ('b' stands for 'beauty' or 'bottom').

The work builds on B-meson observations made by experiments including those at the Large Hadron Collider (LHC) at CERN, Europe's particle-physics laboratory near Geneva, Switzerland. Both efforts are looking for the subtle influence of any new particles or processes on the ways in which known particles decay into others.

Physicists at the LHC have seen some intriguing signs of potential departures from the standard model, most recently in 2017 (The LHCb collaboration *et al. J. High Energ. Phys.* **2017**,

▶

▶ 55; 2017). Buzz around these results has piqued theorists' interest in Belle II, and has prompted new groups to join the international collaboration, says Tom Browder, a physicist at the University of Hawaii at Manoa and spokesperson for the Japan-based experiment.

CLEANER PHYSICS

The collisions at the Belle II experiment will be cleaner and more precise than those at the LHC experiment, called LHCb. That is because the LHCb experiment smashes together protons, which are each composed of three quarks and so make for messy collisions. But Belle II will crash electrons and positrons into each other, both of which are fundamental and so cannot break down any further.

Belle II will be able to study decays involving elusive neutrinos and photons that are harder to investigate with LHCb. This could help it to spot evidence for hypothetical particles, such as charged versions of the Higgs boson — a particle discovered at the LHC in 2012 — and particles such as the axion, a form of dark matter thought to interact with matter only very weakly, says Browder. "There's definitely competition between the two, but also complementarity."

The collider feeding the Belle II experiment will squeeze particles into a tight beam just 50 nanometres across, an advance that will lead to a collision rate 40 times that achieved by its KEK predecessor. This will help it to explore reams of recently discovered exotic particles made up of four or five quarks — tetraquarks and pentaquarks, respectively — and allow it to scour rare b-quark decays for any as-yet unknown preference towards the production of matter over antimatter. It will enable physicists to explore intriguing signs of physics beyond the standard model, a theory that has been verified repeatedly by experiments since the 1970s, but which fails to account for gravity or a host of other mysteries.

Collider experiments produce sprays of many particles that can live for tiny fractions of a second before decaying into other particles. In a handful of decays — involving the transformation of certain B-mesons into electrons and their heavier cousins, called muons and taus — LHCb has seen particles produced at unexpected rates.

It will enable physicists to explore intriguing signs of physics beyond the standard model.

Although each individual finding could easily be a statistical fluctuation, together they have gained attention, says Giovanni Passaleva, a physicist at the National Institute for Nuclear Physics in Florence, Italy, and spokesperson for the LHCb experiment. They broadly point in the same direction and build on similar findings from two previous experiments: the BaBar Collaboration at the SLAC National Accelerator Laboratory in Menlo Park, California; and Belle II's predecessor at KEK, he says. "So it looks like there is some correlation in these deviations, which make them more interesting than others."

SCHEDULED CATCH-UP

However, Belle II will need to catch up with LHCb, whose accelerator produces more B-mesons and has been running since 2009. Once the full physics programme gets under way at the start of 2019, Belle II will take around a year to gather enough data to compete with LHCb. Meanwhile, LHCb will collect data from May until it shuts down for upgrades in November. By then, it should have seen enough decays to either dispel the potential signal or push it into discovery territory. "Our hope is that we get the machine and the detector working fast enough so we can start to catch up with them," says Browder.

The race to claim discovery will come down to which decays prove the most revealing, says Browder. But even if LHCb gets there first, confirmation of new physics from Belle II will be "absolutely essential", says Passaleva. Differences between the two experiments mean that Belle II could help physicists to work out what is behind any new interaction, and definitively rule out experimental error. "Then we'd be sure it's really new physics," he says, "because it will be seen by a completely different experiment in a completely different environment."■

IMMIGRATION

Uncertainty grows for US 'Dreamer' scientists

Court temporarily revives protections against deportation as Congress mulls policy reform.

BY CHRIS WOOLSTON

Like other young researchers in graduate school, Evelyn Valdez-Ward has a lot on her plate. An ecology student at the University of California, Irvine, she has been running field experiments and scrounging for research funding. But, above all, she is worried about whether she can stay in the United States. "My first year has been a real whirlwind," she says. "On top of how difficult grad school is, Trump got elected."

Her future depends on a US government programme that the president, Donald Trump, has attempted to shut down. Known as Deferred Action for Childhood Arrivals (DACA), it shields nearly 800,000 people from deportation, all of whom were brought to the United States illegally as children. Last September, Trump moved to end the programme, prompting a flurry of lawsuits. On 9 January, a federal judge in San Francisco, California, ordered the government to continue DACA while one of the court cases proceeds.

That is little comfort to Valdez-Ward. "If DACA expires, there's no way I can finish my PhD. I would lose everything."

Former president Barack Obama established the DACA programme in 2012 to give young, undocumented immigrants access to legal employment and more forms of financial aid for university studies. To enrol, immigrants must prove that they came to the United States before their sixteenth birthday and have a high-school diploma or are studying for one, among other requirements. Those who are granted DACA status — known as Dreamers —must apply to renew it every two years. Without such protections, they risk being sent back to countries they might not remember, and whose language they might not speak.

Trump's move last year to end DACA prompted lawsuits from 19 states and Washington DC, among other challengers. The case that ultimately led federal judge William Alsup to order DACA's reinstatement was filed by the University of California system — which estimates that some 4,000 of its students are in the country illegally, and that many are probably eligible for DACA status.

"DACA empowered people to start making

DREW ANGERER/GETTY

Many people are pushing for legislation that would give US 'Dreamers' a path to citizenship.

investments in their future, to go to college and medical school," says Roberto Gonzales at Harvard University in Cambridge, Massachusetts, who studies how immigration policies affect the lives of undocumented US immigrants. "Now, that's been thrown into peril."

DACA helped engineering student Josue De Luna Navarro to attend the University of New Mexico in Albuquerque. But he fears that the programme could end. "I remember sitting in a chemical-engineering class trying to calculate a molecule moving through a membrane," he says. "How can I focus on something like that when there's a huge terror in my family and my community about deportation?"

Trump and the US Congress are attempting to negotiate legislation to overhaul US immigration policies — which could end DACA, or shore up the programme. On 11 January, a group of six Democratic and Republican senators announced a compromise that would give DACA recipients a path to citizenship while bolstering border security, but Trump rejected the plan. He has argued that Obama lacked the authority to establish the DACA programme.

Ongoing court cases might determine DACA's short-term future, but its ultimate fate lies with Congress, says Michael Olivas, director of the Institute for Higher Education Law and Governance at the University of Houston in Texas. "This is not a legal issue," he says. "Comprehensive immigration reform, or at least a DACA bill without a bunch of other things attached to it, is the answer." ■

GENETICS

Synthetic species can elude gene mixing

Engineered organisms cannot breed with wild cousins.

BY EWEN CALLAWAY

Maciej Maselko has made wild sex deadly — for genetically modified organisms. The synthetic biologist at the University of Minnesota, Twin Cities, in St Paul and his colleagues have used gene-editing tools to create genetically modified yeasts that cannot breed successfully with their wild counterparts. In so doing, they say they have engineered synthetic species.

"We want something that's going to be identical to the original in every way, except it's just genetically incompatible," says Maselko, who presented his work on 16 January at the annual Plant and Animal Genome Conference in San Diego, California. The research was co-led by Michael Smanski, a biochemist at the University of Minnesota.

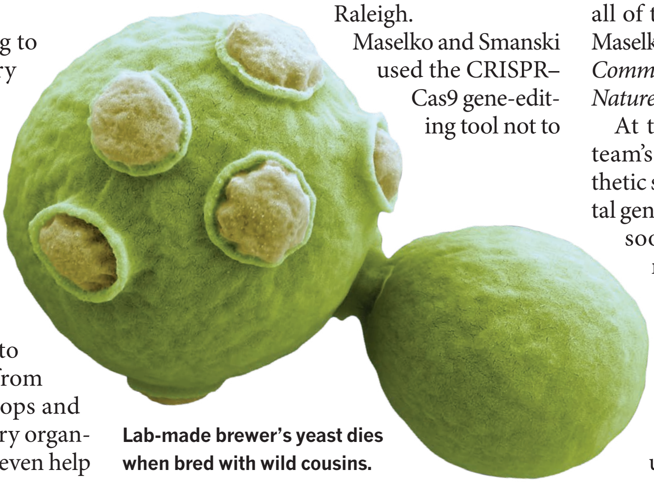
The technology could be used to keep genetically modified plants from spreading genes to unmodified crops and weeds, thereby containing laboratory organisms, the researchers hope. It might even help combat pests and invasive species, by replacing wild organisms with modified counterparts. Other scientists say that the approach is promising, but warn that it could be stymied by technical hurdles, such as the ability of modified organisms to survive and compete in the wild. "This is an ingenious system and, if successful, could have many applications," says evolutionary biologist Fred Gould of the North Carolina State University in Raleigh.

Maselko and Smanski used the CRISPR–Cas9 gene-editing tool not to edit target genes, but to alter their expression. The team guided the Cas9 enzyme to over-activate genes so that their protein products accrued to toxic levels. When they first tested the approach in brewer's yeast (*Saccharomyces cerevisiae*), they raised the levels of a protein called actin to the extent that the cells containing it exploded.

To prevent genetically modified yeast cells from mating successfully with other strains, the team engineered two modifications to the yeast cells. One change was analogous to a 'poison': it produced a version of Cas9 that worked with other factors to recognize and over-activate the actin gene. The second modification, the 'antidote', was a mutation that stopped Cas9 from overexpressing actin.

A yeast strain that contained both poison and antidote produced healthy offspring when mated with a strain carrying the antidote. But when the modified strain was crossed with a different lab strain lacking the antidote, almost all of their offspring popped like balloons, Maselko and Smanski's team reported in *Nature Communications* in October (M. Maselko *et al. Nature Commun.* **8,** 883; 2017).

At the meeting, Maselko discussed the team's progress towards engineering a synthetic species of fruit fly, using a developmental gene called wingless as a poison. Work will soon commence in plants, mosquitoes, nematodes and zebrafish, says Maselko, who, with Smanski, has applied to patent the approach.



EYE OF SCIENCE/SPL

Lab-made brewer's yeast dies when bred with wild cousins.

A COUNTER TO INVASION

A synthetic species could also be used to outcompete and control undesirable species that spread ▶

▶ disease or harm ecosystems. In another contribution to the conference, Maselko's colleague Siba Das, also at the University of Minnesota, presented a mathematical model showing how synthetic speciation could combat invasive carp, which have ravaged rivers and lakes in Minnesota and other central US states.

However, the genetic modifications that stop interbreeding — the poison and antidote — could carry a steep evolutionary fitness cost, says Omar Akbari, a molecular biologist at the University of California, San Diego. The Cas9 enzyme doesn't always recognize its intended gene and could crank up the activity of other genes. Such 'off-target effects' could sap the health of modified organisms. "I'm not sure if this is going to generate a fit-enough strain to compete in the wild," Akbari says.

Gould agrees that it will be difficult to engineer reproductive barriers without incurring evolutionary costs. Scientists could potentially overcome this obstacle by releasing large numbers of modified organisms to increase the odds that a synthetic species will overtake wild organisms. Still, Gould — who is working on other genetic approaches to combating pests — is enthusiastic to see another technology. "I would never want to put all my eggs in one basket," he says. ■

MACHINE LEARNING

Chinese firms enter the battle for AI talent

Country's ambition to become global leader in artificial intelligence needs large workforce.

BY DAVID CYRANOSKI

A mountainous district in western Beijing known for its temples and mushroom production is tipped to become China's hub for industries based on artificial intelligence (AI). Earlier this month, the Chinese government announced that it will spend 13.8 billion yuan (US$2.1 billion) on an AI industrial park — the first major investment in its plan to become a world leader in the field by 2030.

But scientists there wonder whether the proposed 55-hectare AI park, in the Mentougou district 30 kilometres away from the city centre, will be able to attract enough researchers. The government wants it to house 400 companies that will make an estimated 50 billion yuan per year developing products and services in cloud computing, big data, biorecognition and deep learning. "I don't see any top talent willing to go to work and live there," says a scientist working at an AI start-up in Beijing, who asked to remain anonymous because the government is sensitive to criticism.

Sourcing accomplished AI researchers is a problem that's confronting AI-related companies and research centres around the world. "The future [of AI] is going to be a battle for data and for talent," says David Wipf, lead researcher at Microsoft Research in Beijing.

TALENT GRAB

Chinese AI companies are progressing at a dizzying pace. At least five companies developing facial recognition technologies — including SenseTime and Face++, both based in Beijing — pulled in more than $1 billion from investors in 2017. But many AI companies there are struggling to hire researchers. In 2016, the information-technology ministry estimated the country needed an additional 5 million AI workers to meet the industry's needs.

ET大脑像登月计划一样
成为人工智能未来10年重要的研究平台
Like the moon landing program,ET Brain will become the most important platform for artificial intelligence in the next 10 years.

Zhang Yong, head of Chinese tech giant Alibaba, introduces the company's AI, called ET Brain, in 2017.

LI XIN/XINHUA VIA ZUMA

The global pool of experienced AI talent is small. Chinese businesses also have to compete with the aggressive hiring techniques of multinational players such as Google, which some fear are draining universities of researchers by tempting them with high salaries. "It's a talent war — whoever makes the best offer wins," says Nick Zhang, president of the Wuzhen Institute, an AI think tank. He knows of experienced people getting salary offers of $1 million or more to work at the AI research centres of Chinese social-media giant Tencent or the web-services firm Baidu. "This was unimaginable five years ago," he says.

Accomplished industry veterans might be scarce in China, but the country is rich in bright, hard-working computer-science graduates who have expertise in machine learning and other AI-related fields. Peking University in Beijing established the country's first undergraduate course in AI in 2004, and since then 30 universities have introduced similar courses.

But universities are struggling to meet industry's demands, especially because many of the best graduates leave the country. Young Chinese researchers populate AI laboratories from the United States to Israel. At a December 2017 workshop held at New York University (NYU) Shanghai, called Future Leaders of AI Retreat, almost all of the attendees were Chinese researchers working at US universities or industrial laboratories. Zhang Zheng, an AI researcher at NYU Shanghai who organized the retreat, says that he often writes letters of recommendation for Chinese students to study in the United States. “The hope is for them to return later on in their career trajectories,” he says.

There’s also stiff competition for AI researchers within China. Most of the country’s leading AI scientists go to work in industry rather than in academia, says Zhang Zheng. Wipf says that Microsoft set up in Beijing partly to hire the best graduates coming out of nearby Peking and Tsinghua universities, the nation’s premier higher-education institutions.

Last month, Google also established its own AI research centre in Beijing to attract these prodigies. Zhang Zheng says it’s good for the Chinese AI community that international companies are setting up there, because US companies such as Google and Facebook do more fundamental research than local tech giants, he says. “China is lacking top talent, and [working at China-based foreign research hubs] is a way to train them.”

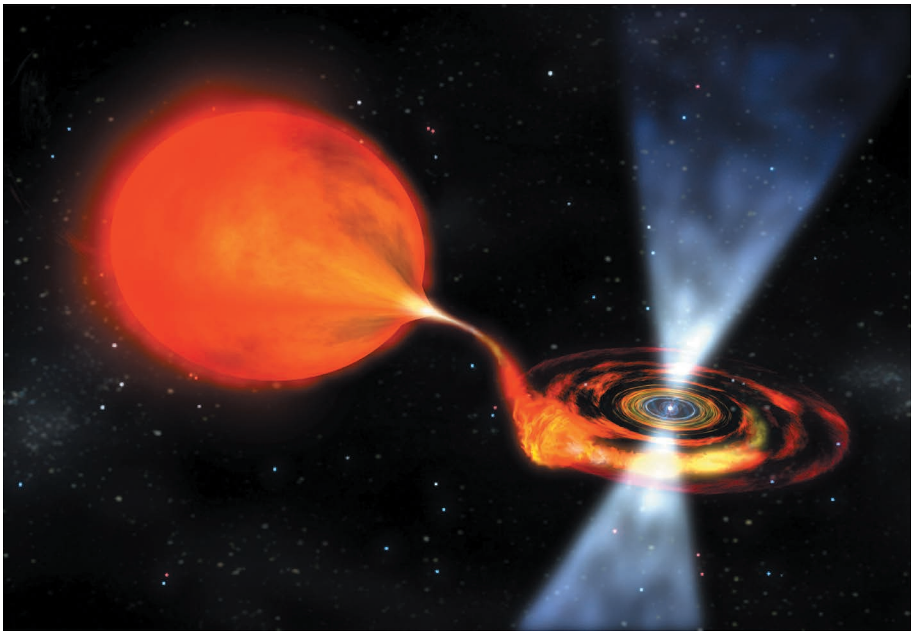
AI TRAINING

The Chinese government realizes that it needs to train and retain more AI graduates if it is to become the world leader in the field by 2030. Its AI road map, released by the Communist Party’s powerful State Council last July, calls for increased education in AI at primary and middle schools.

Online AI training courses are also becoming popular. “The enthusiasm for learning AI is very high,” says Zhang Jiang, who teaches AI at Beijing Normal University’s School of Systems Science.

The country still trails behind the United States in most AI indicators, such as private investment and number of patents, according to figures from the Wuzhen Institute. Nick Zhang says that gap is closing fast, especially in applications such as computer vision.

There’s greater uncertainty about whether China will be able to achieve pioneering breakthroughs in the next decade. “There is still a very big gap before China can lead the competition, because it lacks fundamental innovations,” says Zhang Jiang. “China is still a good learner, but not a good innovator.” ■ SEE EDITORIAL P.249



DANA BERRY/NASA

A pulsar (artist’s impression) gives off beams of radiation as it sucks matter from a companion star.

SPACE SCIENCE

Pulsars can function as a celestial GPS

Experiment shows how spacecraft could use stellar signals to navigate in deep space without human instruction.

BY ALEXANDRA WITZE

From its perch aboard the International Space Station, a NASA experiment has shown how future missions might navigate their way through deep space. Spacecraft could triangulate their location, in a sort of celestial Global Positioning System (GPS), using the regular, rhythmic signals from distant dead stars.

Last November, the Neutron Star Interior Composition Explorer (NICER) spent a day and a half looking at a handful of pulsars — rapidly spinning stellar remnants that give off beams of powerful radiation as they rotate. By measuring tiny changes in the arrival times of the pulses, NICER could pinpoint its location to within 5 kilometres.

It is the first demonstration in space of the long-sought technology known as pulsar navigation. One day, the method could help spacecraft steer themselves without regular instructions from Earth.

“We think it’s a big deal,” said Keith Gendreau, an astrophysicist at NASA’s Goddard Space Flight Center in Greenbelt, Maryland, and the mission’s principal investigator. “It’s a great way to apply some of our astrophysics to exploration goals that include going into the outer Solar System and beyond.” ▶

MORE ONLINE

TOP NEWS

WEIZMANN INST. OF SCIENCE

‘Bat-nav’ reveals how the brain tracks other animals go.nature.com/2mcoaui

MORE NEWS

- University of Rochester president resigns as sexual-harassment probe ends go.nature.com/2dbttpg
- Latest science search engine links papers to grants and patents go.nature.com/2rdn7sx

NATURE PODCAST

Pinning down the climate’s carbon-dioxide sensitivity, and the battle over babies’ first bacteria nature.com/nature/podcast

▶ Gendreau reported the findings on 11 January at a meeting of the American Astronomical Society in National Harbor, Maryland.

The NICER work is a useful test of pulsar navigation under real flight conditions, says John Pye, manager of the Space Research Centre at the University of Leicester, UK, who has worked on the idea.

STELLAR BEACONS

Pulsars are the spinning, ultra-dense leftovers of exploded stars. Some emit radiation blasts as often as every few thousandths of a second. For decades, aerospace engineers have dreamed of using these consistently repeating signals for navigation, just as they use the regular ticking of atomic clocks on satellites for GPS.

In 1999–2000, the US Naval Research Laboratory flew a satellite experiment that showed that, in theory, spacecraft could orient themselves using pulsars. The European Space Agency has explored the concept in recent years, with researchers calculating that a spacecraft could use pulsars to locate itself with a margin of error of 2 kilometres, even when flying 30 times farther from Earth than Earth is from the Sun (S. Shemar *et al. Exp. Astron.* **42,** 101–138; 2016).

In November 2016, China launched an experimental pulsar-navigation satellite, called XPNAV-1. It studied the Crab pulsar, 2,000 parsecs (6,500 light-years) away in the constellation Taurus, as an early test of whether it could lock onto X-ray signals (X. Zhang *et al. Int. J. Aerosp. Eng.* **2017,** 8561830; 2017).

“We think it’s a big deal.”

NICER was installed on the International Space Station in June 2017. Its main job is to measure the size of pulsars to improve scientists’ understanding of the ultra-dense matter that makes up these dead stars. The pulsar-navigation experiment, known as the Station Explorer for X-ray Timing and Navigation Technology (SEXTANT), is a bonus.

SEXTANT timed X-ray flashes coming from five pulsars, one of which is the closest and brightest known millisecond pulsar. The mission watched each of the beacons for about 5–15 minutes before swivelling autonomously to look at the next. By measuring tiny changes in the signals’ arrival times as the experiment orbited Earth, NICER could independently calculate its own position in space.

Without pulsar navigation, spacecraft must communicate with Earth regularly to confirm their position. But such communication — through systems such as NASA’s Deep Space Network, a group of giant satellite dishes — is time-consuming, expensive and more difficult the farther from Earth a probe travels. Pulsar navigation might work well for spacecraft in the outer Solar System because it could free probes to do many navigation-related tasks without waiting for instructions, says Gendreau.

FREEDOM TO ROAM

The technique could also provide an independent check on how well a spacecraft’s conventional navigation systems are doing, says Zaven Arzoumanian, an astrophysicist at Goddard who is on the NICER team. NASA helped fund the test to see whether pulsars could be used as a back-up navigation method when its planned Orion crew capsule takes astronauts beyond low Earth orbit, some time in the 2020s.

NICER used 52 small X-ray telescopes for its study, but a single such telescope could probably do the job, Gendreau says. The instrument might weigh as little as 5 kilograms, making it relatively inexpensive to add to space missions, for which more mass means more money is needed for a launch.

The team plans to repeat the experiment in the coming months, hoping to reduce the margin of error to one kilometre or less. ■

Most infants first come into contact with microbes during birth — or so researchers have assumed.

EDGARD GARRIDO/REUTERS

Baby's first bacteria

THE WOMB WAS THOUGHT TO BE STERILE. SOME SCIENTISTS ARGUE IT'S WHERE THE MICROBIOME BEGINS.

By Cassandra Willyard

Soon after conception, a human embryo begins to assemble a remarkable organ crucial to its survival. The placenta is both a lifeline and a guardian: it shuttles oxygen, nutrients and immune molecules from the mother's bloodstream to her developing fetus, but it also serves as a barrier against infections. For more than a century, doctors have assumed that this ephemeral structure — like the fetus and the womb itself — is sterile, unless something goes wrong.

Starting around 2011, Indira Mysorekar began questioning this idea. She and her colleagues had sliced and stained samples from nearly 200 placentas collected from women giving birth at a hospital in St Louis, Missouri. When the researchers examined the samples under a microscope, they found bacteria in nearly one-third of them[1]. "They were actually inside cells there," says Mysorekar, a microbiologist at Washington University in St Louis.

Bacteria often signal infection, and infections are a common cause of premature birth. But the microbes that Mysorekar observed didn't seem to be pathogens. She didn't see any immune cells near them; nor did she see signs of inflammation. And bacteria weren't present only in the placentas of women who gave birth early; Mysorekar also found them in samples from women who had normal, healthy pregnancies. "That was our first hint that this may be like a normal microbiome," she says.

Studies seeking to understand how microbes help to shape human health and development have become extremely popular over the past few decades, but some researchers are concerned that a crucial question — when bacteria first colonize the body — has not yet been answered. Doctors have assumed that the first contact with colonizing bacteria occurs in the birth canal. Clinicians are even looking to see whether babies born by caesarean section might benefit from a swab of their mother's vaginal microbes. But Mysorekar and other scientists have found evidence of bacteria in the placenta, amniotic fluid and meconium — the tar-like first stool that forms in a fetus *in utero*. This has led some researchers to posit that the microbiome might be seeded before birth.

If that is true and bacteria are a normal — perhaps even crucial — part of pregnancy, they could have an important role in shaping the developing immune system. Scientists might be able to find ways to shift the microbial composition in the womb and possibly ward off allergies, asthma and other conditions. They might also be able to uncover microbial profiles associated with preterm birth or other complications during pregnancy, which could help to illuminate why they occur.

"IF WE DO NOT HAVE MICROBES *IN UTERO*, I THINK WE WOULD BE THE ONLY SPECIES THAT HAS BEEN INTERROGATED THAT DOESN'T."

The scientists at the centre of these discoveries argue that the dogma of a sterile womb is on its way out. Perhaps humans, like species such as clams, tsetse flies and turtles, can inherit a mother's microbes before they are even born[2]. "If we do not have microbes *in utero*, I think we would be the only species that has been interrogated that doesn't," says Susan Lynch, a microbiologist at the University of California, San Francisco.

But even as the number of papers supporting this idea grows, some scientists are pushing back. "I just don't think that these microbiomes exist," says Jens Walter, a microbiologist at the University of Alberta in Edmonton, Canada. Where some see an intriguing new avenue of research, others see biological implausibility, sloppy science and a spectre that has long haunted microbiome research — contamination. Now, studies are getting under way that could answer the question once and for all.

One paediatrician likens the controversy over the placental microbiome to a scientific "knife fight". But if fetal microbiomes do exist, that could have far-reaching implications not only for medicine, but also for basic biology. "If we start thinking of the placenta as a conduit or facilitator of maternal-fetal communication and not as a barrier, then I think we open ourselves up to very interesting perspectives on how we've interpreted a lot of developmental biology today," says Kjersti Aagaard, an obstetrician at Baylor College of Medicine in Houston, Texas.

PROBING THE PLACENTA

The sterile-womb dogma goes back to French paediatrician Henry Tissier, who investigated the source of a baby's first bacteria around the turn of the twentieth century. Researchers began to find bits of evidence against sterility more than three decades ago, but the idea that the placenta might harbour a fully fledged microbiome didn't gain much attention until 2014, when a team of researchers led by Aagaard identified bacterial DNA in placental tissue[3].

Aagaard, who was working on the Human Microbiome Project, noticed something odd. Babies were supposed to get the bacteria that will become their microbiome in the birth canal, but she saw a mismatch between the bacteria present in the vaginas of pregnant women and those present in infants in their first week of life. That might make sense, she thought, if the microbiome gets seeded before birth.

Aagaard reasoned that if mothers were passing bacteria to their babies in the womb, there might be evidence of that transfer in the placenta, the organ that connects the two. To investigate, she and her team harvested tiny bits of tissue under sterile conditions from the placentas of 320 women, including some who gave birth early and some who had infections during pregnancy. Bacteria can be difficult to culture. So, to identify what was there, they used gene sequencing. They took biopsies of the placentas in a sterile room within an hour of delivery, sliced off the surfaces to avoid contamination, and placed those samples into vials. They also analysed the contents of empty vials to rule out contamination from the environment or the DNA-extraction reagents.

Not every placenta contained detectable bacterial DNA, but many did[3]. To get a more in-depth picture of the capabilities of these microbes, the researchers performed whole-genome sequencing on a subset of the samples. In most, they found communities dominated by *Escherichia coli* and a few other groups. And when they compared the bacterial DNA from placentas with that from bacteria typically found in other areas of the body, the results best matched the kinds of microbe found in the mouth. How oral bacteria would have made their way to the placenta isn't clear, but one possibility is that they travelled through the bloodstream. Even routine tooth brushing can allow bacteria access to the blood. What's more, the microbial signature seemed to differ in women who had experienced a preterm birth or an earlier infection. Physicians have assumed that the mere existence of bacteria in the placenta signals infection, but to Aagaard it seemed clear that which bacteria are present is much more important than whether they are there at all.

The paper made a splash in the popular press, but critics argued that Aagaard was overreaching. "DNA is not bacteria," says Mathias Hornef, head of the Institute of Medical Microbiology at the University Hospital RWTH Aachen in Germany. DNA can be used to characterize a microbiome, he says, but not to establish its existence.

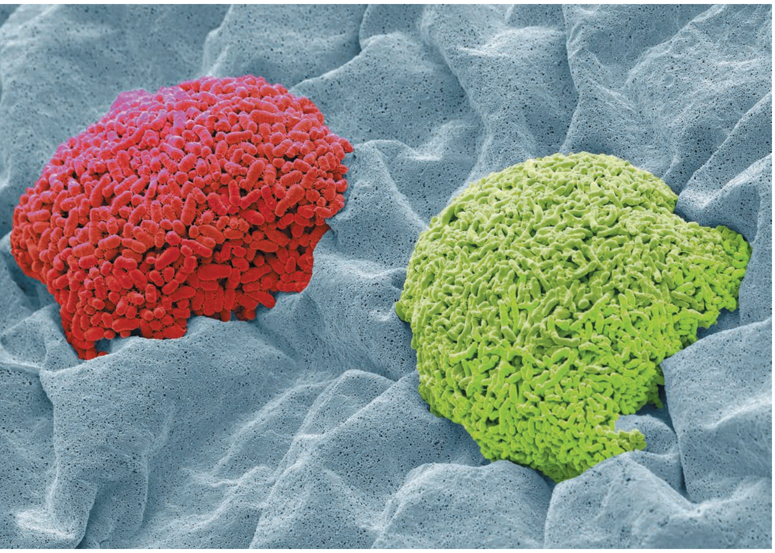
Aagaard's findings weren't an isolated event, however. Several other groups have found bacterial DNA and more in the placenta. Mysorekar, for example, saw the host of bacterial structures inside cells taken from the placenta[1]. And in 2016, a Finnish group managed to culture bacteria from placental tissues taken from women who had healthy pregnancies[4].

Researchers have also found bacteria in amniotic fluid[4,5], leading them to wonder whether the fetus might occasionally ingest microbes when it swallows some of that fluid. And some researchers, including Josef Neu, a neonatologist at the University of Florida in Gainesville, identified bacterial DNA in meconium[6], a finding that suggests the fetus's gut itself may harbour bacteria before birth. Some of the DNA came from the same genera found in amniotic fluid. And the results showed that the microbes in the stool of preterm infants were different from those in babies born at full term.

Neu hypothesized that some strains of bacteria might prompt the fetal gastrointestinal tract to produce inflammatory proteins that would trigger early labour. And indeed, some studies[7] have shown that amniotic fluid from premature babies does hold more of these proteins. That association doesn't prove anything, but it does provide "some interesting pieces of the puzzle", he says. "The fetal–maternal microbiome may be at least a partial explanation for some of these cases of preterm delivery."

Lynch's group is one of several that have been able to culture bacteria from meconium. But it's not yet clear whether those bacteria are simply passing through the fetus, or whether they're actually growing, dividing and taking up residence in the fetal gut, she says. Lynch is now looking at human fetal tissue to see whether she and her colleagues can find evidence of bacteria in the intestinal lining.

A handful of animal studies suggests that this kind of bacterial transfer from mother to fetus is possible. In the mid-2000s, a team of researchers led by microbiologist Juan Miguel Rodríguez at the Complutense University of Madrid inoculated pregnant mice with labelled bacteria, and delivered the pups by caesarean section. They found the



Bacterial culture from a belly button: there is some debate as to how different parts of the body are first seeded with microbes.

STEVE GSCHMEISSNER/SPL

labelled bacteria in both the amniotic fluid[8] and pups' meconium[9].

"What we're seeing in these animal models and what we're seeing in humans really seems to support this fetal–maternal microbiome," says Neu. "I'm not 100% convinced, but I think the data is becoming very strong."

CONTAMINATION QUESTIONS

A number of researchers, however, remain deeply sceptical. The traces of placental microbes, they argue, are 'kitome' — contaminants from the DNA-extraction kits used in the research. There's some evidence to support this. Samuel Parry, a perinatologist at the University of Pennsylvania's Perelman School of Medicine in Philadelphia, was initially intrigued by Aagaard's data. So he planned a study to examine differences between the placental microbiomes of preterm infants and those of babies born at term. As a first step, his team sought out trace amounts of DNA found on sterile swabs, reagents, DNA-purification kits and other equipment that they would routinely use. The bacterial DNA that they ultimately recovered from six placenta samples was indistinguishable from that found on the extraction kits[10]. They've since tested several dozen placentas, Parry says. "We just can't find a microbiome." Marcus de Goffau, a microbiome researcher at the Wellcome Sanger Institute in Hinxton, UK, says that he and his colleagues have similar unpublished results from "hundreds" of placentas.

One of the problems, he says, is that any bacterial signal in the placenta would be weak. In faeces or saliva, there are so many bacteria that it's easy to distinguish the microbiome from background contamination. But when microbes are scarce, a true signal is much harder to pick up. The problem goes much further than studies on human fetuses, he adds: "The entire sequencing field is littered with nonsense."

Aagaard stands by her results. "We are very cautious," she says. "Could we be misinterpreting things? Of course. But we have put in the negative and positive controls every place we can." And she points out that several other groups have found evidence of bacterial DNA in the placenta.

Parry and obstetrician Roberto Romero at the National Institute of Child Health and Human Development in Detroit, Michigan, are planning a multi-centre study to examine the question in even more placentas. They hope to hold a meeting to design the protocol in the next couple of months. If all goes well, they could have an answer as soon as next year, Romero says. They have invited Aagaard to participate, and she says she is willing. "Kjersti Aagaard is an outstanding investigator and she has put forth an idea that is interesting, is important and deserves to be tested," Romero says. "This controversy can be solved."

They aren't the only ones looking for answers. de Goffau is part of a team that has received a £1.6-million (US$2-million) grant from the UK Medical Research Council to examine placental tissue and blood for infectious agents that might be correlated with pregnancy complications. And last year, the US National Institutes of Health announced that it would offer funding for research into the early development of the immune system. The announcement specifically mentioned studies to examine how the fetal microbiome gets seeded and evolves, and how that might impact the brain.

If research fails to detect a microbiome in the womb, that doesn't eliminate the possibility that the fetus might encounter microbes there. "There's very little in and on the human body that could be considered sterile," says Juliette Madan, a neonatologist at Dartmouth–Hitchcock Medical Center in Lebanon, New Hampshire. But a handful of microbes does not necessarily mean there's a complex, thriving microbiome. Madan doesn't expect researchers to find any meaningful sharing of bacteria between mother and fetus.

But de Goffau, one of the most vehement critics of the placenta papers, isn't so sure. He has himself managed to detect bacteria in meconium. "It's not completely sterile. That's pretty clear," he says. Although the evidence isn't complete, he adds, a fetal microbiome is at least possible.

Maria Dominguez-Bello, a microbial ecologist at New York University, runs a study looking at the development of the infant microbiome and the potential benefits of putting babies in contact with their mothers' vaginal microbes after a caesarean section. She doesn't find the reports of bacteria in meconium all that convincing, however. She argues that sterility is broken when the amniotic sac breaks, which leaves plenty of time for bacteria to make their way into the infant's gut. "Labour takes hours, during which the baby is swallowing and rubbing against the walls of the birth canal," she adds. Even if a baby is born by caesarean section, it might take hours or even days for the infant to pass its first stool — a window during which it might acquire bacteria outside the womb.

The most compelling evidence that a fetal microbiome doesn't exist, say Dominguez-Bello and others, is the existence of laboratory mice that are free of bacteria. To create these germ-free rodents, pups are surgically delivered from mothers with normal microbiomes and then raised under sterile conditions. "We've done these experiments, and we've done them for 70 years," Walter says. If just one bacterium were present inside the pup, it would quickly colonize, and the protocol would fail. It would be impossible to complete such experiments.

"I would argue that if you talk with real microbiologists, they wouldn't consider it controversial," says Walter. The question, he adds, has already been answered.

Mysorekar, who is a microbiologist, disagrees. Some people are stuck on the idea that the placental microbiome is "fake news", she says. That, she argues, is a shame. "There are some very exciting questions to address." Humans start to develop a repertoire of immune cells while still in the womb, Mysorekar says, which suggests some sort of microbial exposure. She wonders where these microbes come from and how the exposure occurs. "There's so much to learn," she says. But she isn't surprised by the scepticism. In any emerging field, she says, you'll find "some naysayers, some dirty data, but also a lot of compelling new observations which together push the field forward". ■

Cassandra Willyard *is a freelance journalist based in Madison, Wisconsin.*

1. Stout, M. J. *et al. Am. J. Obstet. Gynecol.* **208,** 226.e1–7 (2013).
2. Funkhouser, L. J. & Bordenstein S. R. *PLoS Biol.* **11,** e1001631 (2013).
3. Aagaard, K. *et al. Sci. Transl. Med.* **6,** 237ra65 (2014).
4. Collado, M. C., Rautava, S., Aakko, J., Isolauri, E. & Salminen, S. *Sci. Rep.* **6,** 23129 (2016).
5. DiGiulio, D. B. *Semin. Fetal Neonatal Med.* **17,** 2–11 (2012).
6. Ardissone, A. N. *et al. PLoS ONE* **9,** e90784 (2014).
7. DiGiulio, D. B. *et al. PLoS ONE* **3,** e3056 (2008).
8. Jiménez, E. *et al. Curr. Microbiol.* **51,** 270–274 (2005).
9. Jiménez, E. *et al. Res. Microbiol.* **159,** 187–193 (2008).
10. Lauder, A. P. *et al. Microbiome* **4,** 29 (2016).

THE DARK SIDE OF LIGHT

BY AISLING IRWIN

The world is lit at night like never before. A clutch of experiments is tracking how ecosystems are faring.

It's a summer night near a forest lake in Germany and something unnatural is going on. Beyond the dark waters lapping at the shores, a faint glow emanates from rings of light hovering above the surface. Nearby, bobbing red torchlights — the least-disruptive part of the visible spectrum — betray the presence of scientists on the shoreline. They are testing what happens when they rob the lake creatures of their night.

This experiment near Berlin is the most ambitious of several projects going on in dark patches of countryside around Europe, set up in the past few years to probe what light pollution is doing to ecosystems. Researchers are growing increasingly concerned about the problem. Although many studies have documented how artificial light harms individual species, the impacts on whole ecosystems and the services they provide, such as crop pollination, is less clear. Several field studies hope to provide answers, by monitoring how plant and animal communities respond to both direct light and the more diffuse unnatural luminance of the night sky, known as skyglow.

Ecologists face challenges such as measuring light accurately and assessing how multiple species behave in response. But early results suggest that light at night is exerting pervasive, long-term stress on ecosystems, from coasts to farmland to urban waterways, many of which are already suffering from other, more well-known forms of pollution. It's an important blind spot, says Steve Long, a plant biologist at the University of Illinois at Urbana–Champaign and editor of the journal *Global Change Biology*. "We know a great deal now about the impacts of rising CO_2," he says. "But how extensive are the impacts of light pollution? We're gambling with our future in what we're doing to the environment."

In the 1950s, Dutch physiologist Frans Verheijen began to study

KAMIEL SPOELSTRA/NIOO-KNAW

In mini-ecosystems in the Netherlands, researchers test the effects of artificial light.

how lights attract animals and interfere with their behaviour. And during the 1970s, more biological observations of the impacts of light started popping up in the literature. But it took two lateral-thinking biogeographers — Catherine Rich, president of the Urban Wildlands Group in Los Angeles, California, and Travis Longcore, now at the University of Southern California in Los Angeles — to see the links between them and organize a conference in 2002, followed by a book, *The Ecological Consequences of Artificial Night Lighting* (Island, 2006), pointing out how far the tendrils of the illuminated night extend.

For the vast majority of organisms — whether human, cockroach or wisp of plankton — the cycle of light and dark is an influential regulator of behaviour. It mediates courtship, reproduction, migration and more. “Since life evolved, Earth has changed dramatically, but there have always been light days and dark nights,” says Christopher Kyba, a physicist at the German Research Centre for Geosciences in Potsdam. “When you change it, you have the worry that it could screw up a lot of things”.

The pace of that change is increasing. Striking images from space over the past two decades reveal the extent to which the night is disappearing. Estimates suggest that more than one-tenth of the planet’s land area experiences artificial light at night[1] — and that rises to 23% if skyglow is included[2]. The extent of artificially lit outdoor areas spread by 2% every year from 2012 to 2016 (ref. 3). An unexpected driver of the trend is the widespread installation of light emitting diodes (LEDs), which are growing in popularity because they are more energy efficient than other bulbs (see page 274). They tend to emit a broad-spectrum white light that includes most of the frequencies important to the natural world.

The trend has had profound impacts on some species; lights are well known to disorient migrating birds and sea turtles, for example. Scientists have also found that disappearing darkness disturbs the behaviour of crickets, moths and bats, and even increases disease transmission in birds.

The most lethal effects are perhaps on insects — vital food sources and pollinators in many ecosystems. An estimate of the effects of street lamps in Germany suggested that the light could wipe out more than 60 billion insects over a single summer[4]. Some insects fly straight into lamps and sizzle; some collapse after circling them for hours.

Fewer studies have examined plants, but those that have suggest that light is disrupting them, too. In a study in the United Kingdom[5], scientists took a 13-year record of the timing of bud opening in trees, and matched it up with satellite imagery of night-time lighting. After controlling for urban heat, they found that artificial lighting was linked with trees bursting their buds more than a week earlier — a magnitude similar to that predicted for 2 °C of global warming. A study of soya-bean farms in Illinois[6] found that the light from adjacent roads and passing cars could be delaying the maturation of crops by up to seven weeks, as well as reducing yield.

ECOSYSTEM EFFECTS

Now, the results of some ambitious experiments are coming in. One of the largest is a field experiment in the Netherlands, where eight locations in nature reserves and dark places host several rows of street lamps. The rows are different colours — green, red, white and a control row turned off — and run from a grassland or heath field into a forest[7]. For six years now, scientists and volunteers have used camera traps to monitor the activity of small mammals; automatic bat detectors to record echolocation calls; mist nets for trapping birds; and nest boxes to assess the timing and success of breeding. Botanists are studying the vegetation underneath the lamps.

The team has found physiological evidence of the detrimental effects of light pollution on the health of wild animals. Songbirds roosting around the white light were restless through the night, slept less and had metabolic changes that could indicate poorer health[8]. The project also looked at how light affects bats, which have had mixed fortunes under the explosion of artificial illumination. Some species, such as the common pipistrelle (*Pipistrellus pipistrellus*), feast on the buffet of insects that they find circling lamps. Other, light-shy, bats have lost habitat and have disappeared from some places. In the Netherlands study, red light had no effect on any of the bat species[9], which means it could be deployed instead of white.

“SINCE LIFE EVOLVED, EARTH HAS CHANGED DRAMATICALLY, BUT THERE HAVE ALWAYS BEEN LIGHT DAYS AND DARK NIGHTS.”

But the experiment has yielded some puzzling findings. Several urban studies had found that artificial light at night triggers songbirds to sing earlier in the day. Because females tend to select early-singing males, the shifted dawn chorus might be affecting which birds get to reproduce. But the team in the Netherlands found no effect on any of 14 songbird species[10]. It’s possible that the lighting was too weak to elicit an effect — it is calibrated to reflect the level on country roads and cycle paths, rather than the glare of an urban park.

Both kinds of result are useful for local governments, says Kamiel Spoelstra, a biologist at the Netherlands Institute of Ecology (NIOO-KNAW) in Wageningen, who leads the project. His team’s findings are being incorporated into Dutch regulations on outdoor lighting. For instance, he says, some areas seeking to support local bat populations have switched to red light, a trend that he expects to increase.

Coloured light also graces grasslands in southwest England, where a project known as Ecolight is looking for evidence of ‘cascade effects’, in which the influences of light on one species have knock-on effects on the ecosystem.

The glowing cubes used by Ecolight might be mistaken for an art installation. Scientists led by Kevin Gaston, a biodiversity and

KAMIEL SPOELSTRA/NIOO-KNAW

This grassland experiment supports the idea that red light is relatively benign to wildlife.

conservation specialist at the University of Exeter, UK, have just finished researching 54 artificial communities of grassland. In some of the cubes, beetles, slugs, pea aphids and 18 species of plant muddled along for 5 years, isolated from the outside world. Other boxes were simpler — containing just plants and herbivores, or plants alone. At night, some were illuminated with white light, some with amber, and some just saw the raw sky.

The effects of light on grasslands are important, partly because roadside grass provides refuges and corridors for wildlife in built-up areas. The scientists discovered that amber light and, to a lesser extent, white, suppressed flowering in the trefoil (*Lotus pedunculatus*)[11]. And there was a cascade effect in the amber-lit boxes. During August, when pea aphids switch from eating shoots to feasting on flower heads, their numbers fell, presumably because their food was less abundant. "I think this is the first experimental evidence of a strong, bottom-up effect of exposure to artificial light," says Gaston. In its latest, unpublished, work, the team reveals further effects, cascading onto the predators in the systems.

> **"WE WERE TOO IGNORANT AS BIOLOGISTS ABOUT THE COMPLEXITY OF LIGHT."**

Another elaborate experiment, in a dark-skies reserve in Westhavelland Nature Park in Germany, has shown that these cascade effects can spill over into neighbouring ecosystems. Street lamps erected near water-filled ditches lure aquatic insects out of the water[12], says Franz Hölker, an ecohydrologist at the Leibniz Institute of Freshwater Ecology and Inland Fisheries in Berlin. The insects flock to the lamps, exhaust themselves and become food for nearby predators. Meanwhile, the hinterland, which might otherwise have received insect visits, is deprived of an important source of food, he says.

Studies such as these, which lay such relationships bare in well-controlled, small-scale studies, mean that "those impacts are more likely to be taken seriously in the field and by regulators considering impacts from lighting", says Longcore.

Artificial light can also have impacts on ecosystem services — the benefits that ecosystems provide to humans. A study published in *Nature* last year found that illuminating a set of Swiss meadows stopped nocturnal insects pollinating plants[13]. A team led by Eva Knop of the Institute of Ecology and Evolution at the University of Berne, found that insect visits to the plants dropped by nearly two-thirds under artificial light and that daytime pollination couldn't compensate: the plants produced 13% less fruit. Knop's team forecast that these changes had the potential to cascade to the daytime pollinator community by reducing the amount of food available. "This is a very important study, which clearly demonstrates that artificial light at night is a threat to pollination," says Hölker.

LIGHT SKIES

Much of Earth remains free of direct artificial light, but skyglow — light that is scattered back to Earth by aerosols and clouds — is more widespread. It can be so faint that humans can't see it, but researchers say it could still threaten the 30% of vertebrates and 60% of invertebrates that are nocturnal and exquisitely sensitive to light.

Skyglow "almost certainly" has an impact on biodiversity, Gaston says, because the level is well above the thresholds for triggering many biological responses. And yet, he says, "it's actually quite hard to do the definitive study".

That's where the forest-lake experiment comes in. Glowing circles of light hover above cylinders sunk into Lake Stechlin, recreating skyglow. They are the work of Leibniz physicist Andreas Jechow, who had to find a way to produce low-level, even illumination without blocking daylight or impeding access for scientists. He and his team achieved this using state-of-the-art photonics tools such as an advanced ray-tracing model. "We were too ignorant as biologists about the complexity of light as a physical phenomenon," says Mark Gessner, director of the project, known as The Lakelab, and co-leader of its artificial-light project, called ILES (Illuminating Lake Ecosystems). In the past, some experiments have even failed to account for the fact that the Moon moves across the sky, he adds.

The idea for ILES was to extend findings from a well-known study of zooplankton, which live in deep, dark water during the day and migrate up into shallower waters at night to graze on algae. This movement is thought to be the biggest migration of biomass in the world. A study[14] in lakes near Boston, Massachusetts, in the late 1990s suggested that skyglow reduces the zooplankton's ascent by 2 metres, and the number of organisms that ascend by 10–20%. This behavioural change may be an unacknowledged driver of fundamental lake processes such as algal blooms.

At ILES, the 24 cylinders — each 9 metres in diameter — look from the surface like a fish farm. Lighting them with different levels of 'skyglow' and measuring the distribution of the tiny plankton using video cameras, the scientists found that skyglow had no massive effect on the movement of algae. "We may have a changed migration pattern but I'm not yet certain about this," says Gessner. "If there is an effect, though, it looks like it's not the profound one we were expecting."

The surprise result is typical of these difficult studies. Gessner points out that their experiment has only completed its first season. "Maybe we don't need to be worried or maybe we need to be less worried — we don't know, at least as far as the effects of skyglow on lakes is concerned," he says.

BRIGHT FUTURE

It's slow, meticulous work, but the field is coalescing as evidence accumulates, says Gaston. "The last two or three years has seen a dramatic improvement in the level of our understanding," he says.

Nonetheless, there are improvements to make. Even measuring exposure is hard. In the field, the light an organism receives can be difficult to measure; a bird could retreat to the shadow of a nearby tree to avoid illumination, for example. So some scientists have tried strapping light meters to birds to get a better idea of dosage.

As the results seep out, one thing that both frustrates and inspires ecologists is that the remedy is at hand.

Longcore is now gathering published data on how different species, such as shearwaters and sea turtles, respond to different parts of the spectrum, and matching the results to the spectra emitted by different types of lighting. He wants to inform decisions about lighting — for example, which type of lamp to use on a bridge and which at a seaside resort.

Engineers and ecologists know that well-considered lighting can perform its task without "spraying light into the sky", as Kyba puts it. LEDs can be tweaked to shine in certain parts of the spectrum, to dim and to switch off remotely. "My vision," says Kyba, "is that in 30 years' time, the streets will be nicely lit — better than today — but we'll use one-tenth of the light."

That would be great news for ecological systems, says Hölker, because darkness is one of the most profound forces to shape nature. "Half of the globe is always dark," he says. "The night is half the story." ■

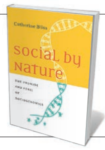
Aisling Irwin *is a science journalist based in Oxford, UK.*

1. Gaston, K. J., Duffy, J. P., Gaston, S., Bennie, J. & Davies, T. W. *Oecologia* **176,** 917–931 (2014).
2. Falchi, F. *et al. Sci. Adv.* **2,** e1600377 (2016).
3. Kyba, C. C. M. *et al. Sci. Adv.* **3,** e1701528 (2017).
4. Eisenbeis, G. in *Ecological Consequences of Artificial Night Lighting*. Rich, C. & Longcore, T. (eds). pp 281–304 (Island, 2006).
5. ffrench-Constant, R. H. *et al. Proc. R. Soc. B* **283,** 20160813 (2016).
6. Palmer, M. *et al.* 'Roadway lighting's impact on altering soybean growth' (Illinois Center for Transportation, 2017); available at http://go.nature.com/2qh3fjb
7. Spoelstra, K. *et al. Phil. Trans. R. Soc. B* **370,** 20140129 (2015).
8. Ouyang, J. Q. *et al. Glob. Change Biol.* **23,** 4987–4994 (2017).
9. Spoelstra, K. *et al. Phil. Trans. R. Soc. B* **284,** 20170075 (2017).
10. Da Silva, A. *et al. R. Soc. Open Sci.* **4,** 160638 (2017).
11. Bennie, J. *et al. Phil. Trans. R. Soc. B* **370,** 20140131 (2015).
12. Manfrin, A. *et al. Front. Environ. Sci.* **5,** 61 (2017).
13. Knop, E. *et al. Nature* **548,** 206–209 (2017).
14. Moore, M. V., Pierce, S. M., Walsh, H. M., Kvalvik, S. K. & Lim, J. D. *SIL Proc, 1922–2010* **27,** 779–782 (2000).

COMMENT

HEALTH Poor artificial lighting puts people, plants and animals at risk **p.274**

GENOMICS Sociology of genetics research reveals baked-in bias **p.278**



AI Design robots to self-certify as safe for autonomous work **p.281**

OBITUARY Ben Barres, glia neuroscientist and equality advocate, remembered **p.282**

KARL NICKEL/OBERWOLFACH PHOTO COLLECTION

Olga Ladyzhenskaya was on the Fields Medal shortlist in 1958.

The Fields Medal should return to its roots

Forgotten records of mathematics' best-known prize hold lessons for the future of the discipline, argues historian **Michael Barany**.

Like Olympic medals and World Cup trophies, the best-known prizes in mathematics come around only every four years. Already, maths departments around the world are buzzing with speculation: 2018 is a Fields Medal year.

While looking forward to this year's announcement, I've been looking backwards with an even keener interest. In long-overlooked archives, I've found details of turning points in the medal's past that, in my view, hold lessons for those deliberating whom to recognize in August at the 2018 International Congress of Mathematicians in Rio de Janeiro in Brazil, and beyond.

Since the late 1960s, the Fields Medal has been popularly compared to the Nobel prize, which has no category for mathematics[1]. In fact, the two are very different in their procedures, criteria, remuneration and much else. Notably, the Nobel is typically given to senior figures, often decades after the contribution being honoured. By contrast, Fields medallists are at an age at which, in most sciences, a promising career would just be taking off.

This idea of giving a top prize to rising stars who — by brilliance, luck and circumstance — happen to have made a major mark when relatively young is an accident of history. It is not a reflection of any special connection between maths and youth — a myth unsupported by the data[2,3]. As some mathematicians have long recognized[4], this accident has been to mathematics' detriment. It reinforces biases within the discipline and in the public's attitudes about mathematicians' work, career pathways and intellectual and social values. All 56 winners so far have been phenomenal mathematicians, but such biases have contributed to 55 of them being male, most being from the United States and Europe and most working on a collection of research topics that are arguably unrepresentative of the discipline as a whole.

When it began in the 1930s, the Fields Medal had very different goals. It was rooted more in smoothing over international conflict than in celebrating outstanding scholars. In fact, early committees deliberately avoided trying to identify the best young mathematicians and sought to promote relatively unrecognized individuals. As I demonstrate here, they used the medal to shape their discipline's future, not just to judge its past and present.

▶

As the mathematics profession grew and spread, the number of mathematicians and the variety of their settings made it harder to agree on who met the vague standard of being promising, but not a star. In 1966, the Fields Medal committee opted for the current compromise of considering all mathematicians under the age of 40. Instead of celebrity being a disqualification, it became almost a prerequisite.

I think that the Fields Medal should return to its roots. Advanced mathematics shapes our world in more ways than ever, the discipline is larger and more diverse, and its demographic issues and institutional challenges are more urgent. The Fields Medal plays a big part in defining what and who matters in mathematics.

The committee should leverage this role by awarding medals on the basis of what mathematics can and should be, not just what happens to rise fastest and shine brightest under entrenched norms and structures. By challenging themselves to ask every four years which unrecognized mathematics and mathematicians deserve a spotlight, the prizegivers could assume a more active responsibility for their discipline's future.

BORN OF CONFLICT

The Fields Medal emerged from a time of deep conflict in international mathematics that shaped the conceptions of its purpose. Its chief proponent was John Charles Fields, a Canadian mathematician who spent his early career in a *fin de siècle* European mathematical community that was just beginning to conceive of the field as an international endeavour[5].

The first International Congress of Mathematicians (ICM) took place in 1897 in Zurich, Switzerland, followed by ICMs in Paris in 1900, Heidelberg in Germany in 1904, Rome in 1908 and Cambridge, UK, in 1912. The First World War derailed plans for a 1916 ICM in Stockholm, and threw mathematicians into turmoil.

When the dust settled, aggrieved researchers from France and Belgium took the reins and insisted that Germans and their wartime allies had no part in new international endeavours, congresses or otherwise. They planned the first postwar meeting for 1920 in Strasbourg, a city just repatriated to France after half a century of German rule.

In Strasbourg, the US delegation won the right to host the next ICM, but when its members returned home to start fundraising, they found that the rule of German exclusion dissuaded many potential supporters. Fields took the chance to bring the ICM to Canada instead. In terms of international participation, the 1924 Toronto congress was disastrous, but it finished with a modest financial surplus. The idea for an international medal emerged in the organizers' discussions, years later, over what to do with these leftover funds.

Fields forced the issue from his deathbed in 1932, endowing two medals to be awarded at each ICM. The 1932 ICM in Zurich appointed a committee to select the 1936 medallists, but left no instructions as to how the group should proceed. Instead, early committees were guided by a memorandum that Fields wrote shortly before his death, titled 'International Medals for Outstanding Discoveries in Mathematics'.

Most of the memorandum is procedural: how to handle the funds, appoint a committee, communicate its decision, design the medal and so on. In fact, Fields wrote, the committee "should be left as free as possible" to decide winners. To minimize national rivalry, Fields stipulated that the medal should not be named after any person or place, and never intended for it to be named after himself. His most famous instruction, later used to justify an age limit, was that the awards should be both "in recognition of work already done" and "an encouragement for further achievement". But in context, this instruction had a different purpose: "to avoid invidious comparisons" among factious national groups over who deserved to win.

"Fields stipulated that the medal should not be named after any person or place."

The first medals were awarded in 1936, to mathematicians Lars Ahlfors from Finland and Jesse Douglas from the United States. The Second World War delayed the next medals until 1950. They have been given every four years since.

BLOOD AND TEARS

The Fields Medal selection process is supposed to be secret, but mathematicians are human. They gossip and, luckily for historians, occasionally neglect to guard confidential documents. Especially for the early years of the Fields Medal, before the International Mathematical Union became more formally involved in the process, such ephemera may well be the only extant records.

One of the 1936 medallists, Ahlfors, served on the committee to select the 1950 winners. His copy of the committee's correspondence made its way into a mass of documents connected with the 1950 ICM, largely hosted by Ahlfors's department at Harvard University in Cambridge, Massachusetts; these are now in the university's archives.

The 1950 Fields Medal committee had broad international membership. Its chair, Harald Bohr (younger brother of the physicist Niels), was based in Denmark. Other members hailed from Cambridge, UK, Princeton in New Jersey, Paris, Warsaw and Bombay. They communicated mostly through letters sent to Bohr, who summarized the key points in letters he sent back. The committee conducted most of these exchanges in the second half of 1949, agreeing on the two winners that December.

The letters suggest that Bohr entered the process with a strong opinion about who should win one of the medals: the French mathematician Laurent Schwartz, who had blown Bohr away with an exciting new theory at a 1947 conference[6]. The Second World War meant that Schwartz's career had got off to an especially rocky start: he was Jewish and a Trotskyist, and spent part of the French Vichy regime in hiding using a false name. His long-awaited textbook had still not appeared by the end of 1949, and there were few major new results to show.

Bohr saw in Schwartz a charismatic leader of mathematics who could offer new connections between pure and applied fields. Schwartz's theory did not have quite the revolutionary effects Bohr predicted, but, by promoting it with a Fields Medal, Bohr made a decisive intervention oriented towards his discipline's future.

The best way to ensure that Schwartz won, Bohr determined, was to ally with Marston Morse of the Institute for Advanced Study in Princeton, who in turn was promoting his Norwegian colleague, Atle Selberg. The path to convincing the rest of the committee was not straightforward, and their debates reveal a great deal about how the members thought about the Fields Medal.

Committee members started talking about criteria such as age and fields of study, even before suggesting nominees. Most thought that focusing on specific branches of mathematics was inadvisable. They entertained a range of potential age considerations, from an upper limit of 30 to a general principle that nominees should have made their mark in mathematics some time since the previous ICM in 1936. Bohr cryptically suggested that a cut-off of 42 "would be a rather natural limit of age".

By the time the first set of nominees was in, Bohr's cut-off seemed a lot less arbitrary. It became clear that the leading threat to Bohr's designs for Schwartz was another French mathematician, André Weil, who turned 43 in May 1949. Everyone, Bohr and Morse included, agreed that Weil was the more accomplished mathematician. But Bohr used the question of age to try to ensure that he didn't win.

As chair, Bohr had some control over the narrative, frequently alluding to members' views that "young" mathematicians should be favoured while framing Schwartz as the prime example of youth. He asserted that Weil was already "too generally recognized" and drew attention to Ahlfors's contention that to give a medal to Weil would be "maybe even disastrous" because "it would make the impression that the Committee has tried to designate the greatest mathematical genius."

Their primary objective was to avoid international conflict and invidious comparisons. If they could deny having tried to select the best, they couldn't be accused of having snubbed someone better.

But Weil wouldn't go away. Committee member Damodar Kosambi thought it would be "ridiculous" to deny him a medal — a comment Bohr gossiped about to a Danish colleague but did not share with the committee. Member William Hodge worried "whether we might be shirking our duty" if Weil did not win. Even Ahlfors argued that they should expand the award to four recipients so that they could include Weil. Bohr wrote again to his Danish confidant that "it will require blood and tears" to seal the deal for Schwartz and Selberg.

Bohr prevailed by cutting the debate short. He argued that Weil would open a floodgate to considering prominent older mathematicians, and asked for an up or down vote on the pair of Schwartz and Selberg. Finally, at the awards ceremony at the 1950 ICM, Bohr praised Schwartz for being recognized and eagerly followed by a younger generation of mathematicians — the very attributes he had used to exclude Weil.

FURTHER ENCOURAGEMENT

Another file from the Harvard archives shows that the 1950 deliberations reflected broader attitudes towards the medal, not just one zealous chair's tactics. Harvard mathematician Oscar Zariski kept a selection of letters from his service on the 1958 committee in his private collection.

Zariski's committee was chaired by mathematician Heinz Hopf of the Swiss Federal Institute of Technology in Zurich. Its first round of nominations produced 38 names. Friedrich Hirzebruch was the clear favourite, proposed by five of the committee members.

Hopf began by crossing off the list the two oldest nominees, Lars Gårding and Lipman Bers. His next move proved that it was not age per se that was the real disqualifying factor, but prior recognition: he ruled out Hirzebruch and one other who, having recently taken up professorships at prestigious institutions, "did not need further encouragement". Nobody on the committee seems to have batted an eyelid.

Of those remaining, the committee agreed that Alexander Grothendieck was the most talented, but few of his results were published and they considered him a shoo-in for 1962. John Nash, born in the same year as Grothendieck (1928), came third in the final ballot. Although the 1958 shortlist also included Olga Ladyzhenskaya and Harish-Chandra, it would take until 2014 for the Fields Medals to go to a woman (Maryam Mirzakhani) or a mathematician of Indian descent (Manjul Bhargava). Ultimately, the 1958 awards went to Klaus Roth and René Thom, both of whom the committee considered promising but not too accomplished — unlikely to provoke invidious comparisons.

MFO/OBERWOLFACH PHOTO COLLECTION

The nomination of French mathematician André Weil divided the 1950 Fields Medal committee.

A SWEEPING EXPEDIENT

By 1966, the adjudication of which young mathematicians were good but not too good had become testing. That year, committee chair Georges de Rham adopted a firm age limit of 40, the smallest round number that covered the ages of all the previous Fields recipients.

Suddenly, mathematicians who would previously have been considered too accomplished were eligible. Grothendieck, presumably ruled out as too well-known in 1962, was offered the medal in 1966, but boycotted its presentation for political reasons.

The 1966 cohort contained another politically active mathematician, Stephen Smale. He went to accept his medal in Moscow rather than testify before the US House Un-American Activities Committee about his activism against the Vietnam War. Colleagues' efforts to defend the move were repeated across major media outlets, and the 'Nobel prize of mathematics' moniker was born.

This coincidence — comparing the Fields Medal to a higher-profile prize at the same time that a rule change allowed the medallists to be much more advanced — had a lasting impact in mathematics and on the award's public image. It radically rewrote the medal's purpose, divorcing it from the original goal of international reconciliation and embracing precisely the kinds of judgement Fields thought would only reinforce rivalry.

Any method of singling out a handful of honorees from a vast discipline will have shortcomings and controversies. Social and structural circumstances affect who has the opportunity to advance in the discipline at all stages, from primary school to the professoriate. Selection committees themselves need to be diverse and attuned to the complex values and roles of mathematics in society.

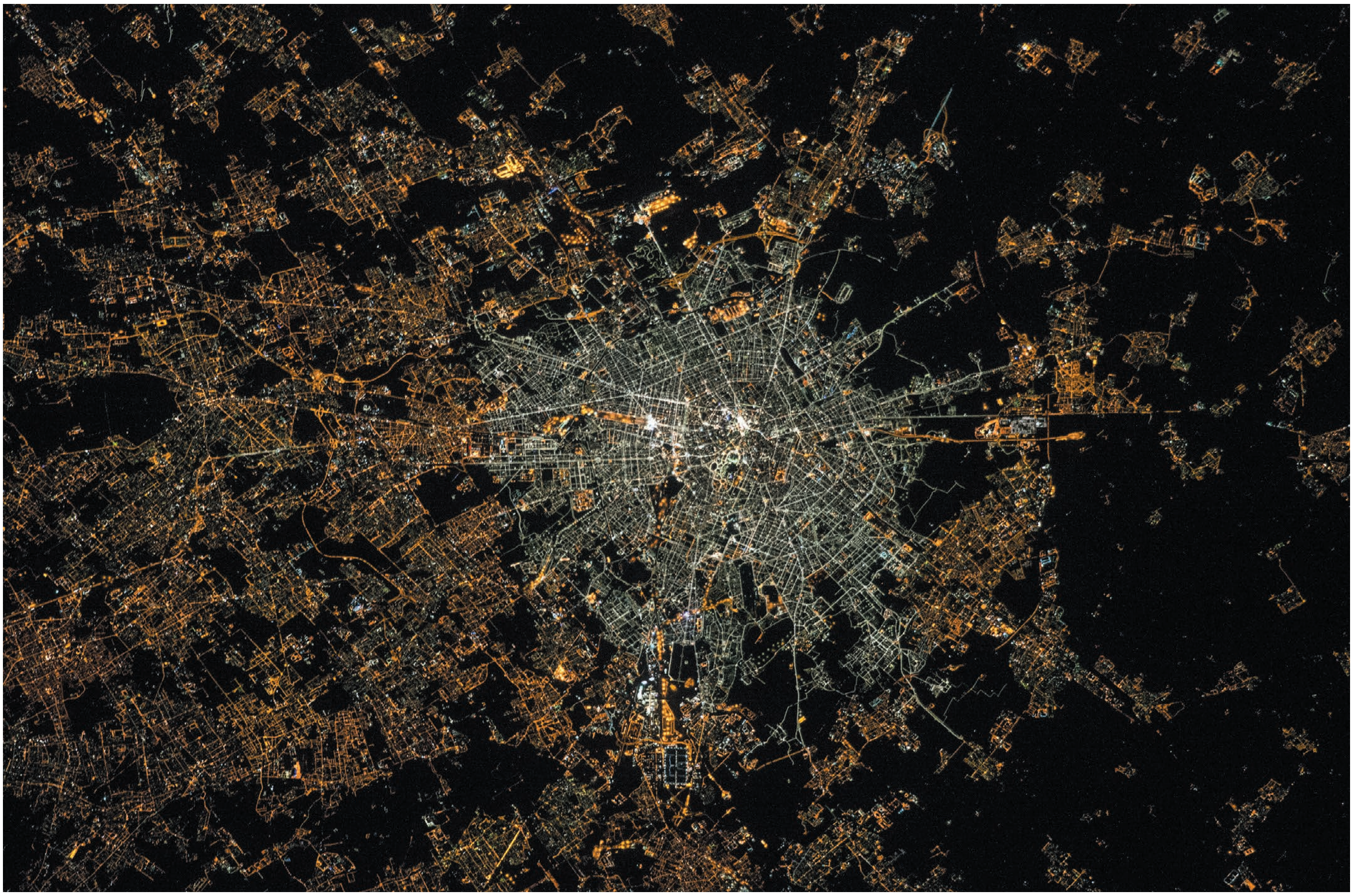
But, however flawed the processes were before 1966, they forced a committee of elite mathematicians to think hard about their discipline's future. The committees used the medal as a redistributive tool, to give a boost to those who they felt did not already have every advantage but were doing important work nonetheless.

Our current understanding of the social impact of mathematics and of barriers to diversity within it is decidedly different to that of mathematicians in the mid-twentieth century. If committees today were given the same licence to define the award that early committees enjoyed, they could focus on mathematicians who have backgrounds and identities that are under-represented in the discipline's elite. They could promote areas of study on the basis of the good they do in the world, beyond just the difficult theorems they produce.

In my view, the medal's history is an invitation for mathematicians today to think creatively about the future, and about what they could say collectively with their most famous award. ■

Michael Barany *is a postdoctoral fellow in the Society of Fellows and Department of History, Dartmouth College, Hanover, New Hampshire, USA.*
e-mail: michael@mbarany.com

1. Barany, M. J. *Not. Am. Math. Soc.* **62,** 15–20 (2015).
2. Stern, N. *Soc. Stud. Sci.* **8,** 127–140 (1978).
3. Hersh, R. & John-Steiner, V. *Loving + Hating Mathematics: Challenging the Myths of Mathematical Life* 251–272 (Princeton Univ. Press, 2011).
4. Henrion, C. *AWM Newsletter* **25** (6), 12–16 (1995).
5. Riehm, E. M. & Hoffman, F. *Turbulent Times in Mathematics: The Life of J.C. Fields and the History of the Fields Medal* (American Mathematical Society & Fields Institute, 2011).
6. Barany, M. J, Paumier, A.-S. & Lützen, J. *Hist. Math.* **44,** 367–394 (2017).



NASA/ESA/SAMANTHA CRISTOFORETTI

Milan in Italy replaced sodium street lighting with blue-rich white LED sources. City-centre illumination now looks brighter and bluer than in the suburbs.

Make lighting healthier

Artificial illumination can stop us sleeping and make us ill. We need fresh strategies and technologies, argues **Karolina M. Zielinska-Dabkowska**.

Life on Earth evolved in day-and-night cycles. Plants and animals, including insects such as the fruit fly, have a biological clock that controls their circadian rhythms — as the 2017 winners of the Nobel Prize in Physiology or Medicine showed. Now, humans' increasing reliance on artificial lighting is changing those rhythms[1].

For more than a century, incandescent light sources served us well. These bulbs were cheap to produce and dispose of, and easy to dim. Their spectrum is continuous and includes most of the colours of the rainbow, much like a sunset (see 'Light-source spectra'). They had their problems. In the 1990s, some researchers blamed electric illumination for changing our sleeping patterns from the natural rhythm of two four-hour phases broken by an hour of wakefulness, to a single eight-hour phase each night. Incandescent lamps are energy hungry and policymakers worried about their contribution to global warming. In 2005, lighting consumed around one-fifth of the world's energy.

In 2009, the European Commission began to withdraw incandescent lamps from the European market. Other countries followed, from Switzerland and Australia to Russia, the United States and China. Low-energy lamps — at first mainly compact fluorescent lamps (CFLs) and later light-emitting diodes (LEDs) — have been promoted as replacements. The health risks this policy poses to humans, animals and plants have yet to be thoroughly assessed.

As a lighting researcher and designer, I am convinced that the costs of this transition far outweigh the benefits for human health and the environment. Because the world's urban population spends more time indoors under artificial lighting than in daylight, the health impacts are already evident. Around one billion people globally lack vitamin D or do not have enough[2]. Seasonal affective disorder, a type of depression that can occur in winter when there is less natural daylight, is on the rise. Shift workers face increased risks of cancer[3], obesity[4] and sleep problems[5].

Biologically benign forms of energy-efficient lighting are needed. I call on physicists, engineers, medical experts, biologists and designers to develop them. Policymakers, planners and regulators should rethink standards, encourage the use of natural light and minimize the negative impacts of artificial lighting at night, indoors and out.

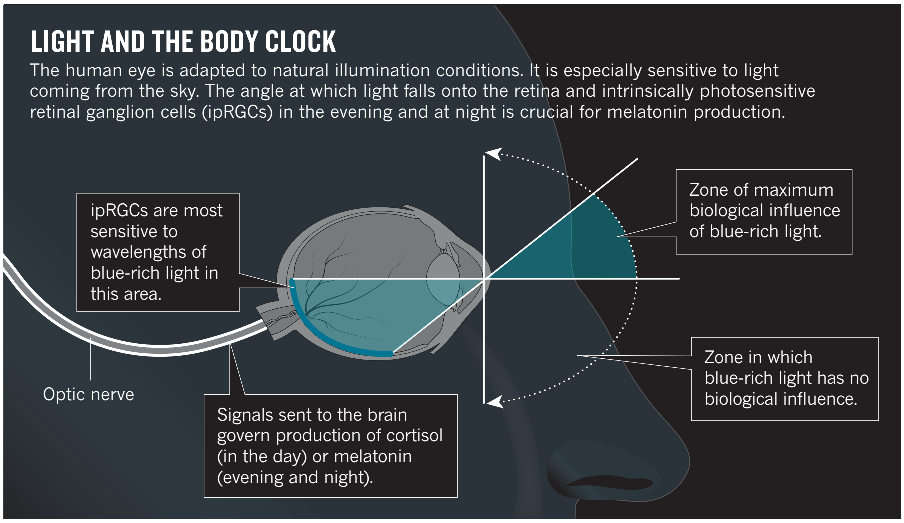
SPIKY SPECTRA

In my view, there is now enough evidence to conclude that the first wave of low-energy light sources is harmful. CFLs are most hazardous. They contain mercury, a neurotoxin. There are no protocols for recycling or disposing of them — 80% are thrown into landfill. Ultraviolet light can escape from defective tube coatings to burn skin or damage the retina at close range; the US Food and Drug Administration recommends coming no nearer than 30 centimetres to a CFL for more than an hour a day.

CFLs have 'spiky' rather than smooth spectra: they emit only certain blue, green and orange-red frequencies (see 'Light-source spectra'). Their flickering at 100–120 hertz can cause headaches and eye fatigue[6]. The energy savings may be overestimated — CFLs take minutes to warm up, so are likely to be left on for longer. When switched on and off many times, they fail more quickly.

Solid-state lighting in the form of LEDs is more promising. LEDs do not contain mercury and produce only a small amount of UV (compared to CFLs or even incandescent lamps). They are more energy efficient, brighter and more long-lived than CFLs. Unlike CFLs, they can be dimmed or tuned and render colours well. But LEDs have downsides[7]. Some contain heavy metals such as nickel, lead and copper, and poisons such as arsenic. Again, there are no special programmes for recycling or disposing of them. Poor-quality LEDs can also flicker and produce stroboscopic effects, such as trails of lights that can confuse pedestrians, cyclists or car drivers.

The lighting industry is beginning to address the lack of daylight in indoor spaces. In recent years, it has promoted artificial, biologically effective lighting in office and home environments, known as human-centric or circadian lighting. This promises to adjust people's daily rhythms in indoor spaces, using LED colour-changing lights that mimic daylight according to the time of the day. The German Commission for Occupational Health and Safety and Standardization (KAN) has issued concerns regarding these practices. The risks of adverse effects remain, because there is still too little understanding of the link between light stimuli and non-visual responses. Research is needed to find out more and to firm up standards accordingly.


LIGHT AND THE BODY CLOCK

The human eye is adapted to natural illumination conditions. It is especially sensitive to light coming from the sky. The angle at which light falls onto the retina and intrinsically photosensitive retinal ganglion cells (ipRGCs) in the evening and at night is crucial for melatonin production.

ipRGCs are most sensitive to wavelengths of blue-rich light in this area.
Zone of maximum biological influence of blue-rich light.
Zone in which blue-rich light has no biological influence.
Optic nerve
Signals sent to the brain govern production of cortisol (in the day) or melatonin (evening and night).

CLAIRE WELSH/NATURE

BLUE PROBLEM

In the meantime, artificial lighting is in my view becoming a public-health hazard. CFLs and LEDs emit more blue light of short wavelengths than a sunset or an incandescent lamp does (see 'Light-source spectra'). Most white LED lamps are made by coating blue or sometimes violet LEDs with yellow pigment, usually phosphor.

The human circadian system is exquisitely sensitive to the spectrum of light visible to the eye, especially blue wavelengths, and its amount and intensity (see 'Light and the body clock'). As well as rod and cone receptors used for vision, the eye contains cells called intrinsically photosensitive retinal ganglion cells (ipRGCs). These send signals to the brain that trigger the body to produce or inhibit neurotransmitters and hormones throughout the day[8]. The spectral sensitivity of melanopsin, the photopigment of ipRGCs, reaches maximum absorbance at approximately 480 nanometres, matching the colour of a clear blue sky at noon.

In the morning, waking is helped by blue wavelengths of daylight triggering releases of the neurotransmitters serotonin and dopamine and the hormone cortisol. In the evening, as natural levels of blue light drop and are replaced by dim red light, melatonin hormone is produced and helps us to fall asleep. Complete darkness is needed at night to initiate processes of cell renewal.

> *"Healthy lighting design is becoming an important ethical issue that cannot be ignored."*

When people are subjected to artificial blue-rich white light at night, from screens and electronic devices as well as artificial illumination, the photosensitive ganglion cells in the retina signal the brain to stop producing melatonin. Such disturbances can have wide effects: on sleep and waking cycles, eating patterns, metabolism, reproduction, mental alertness, blood pressure and heart rate, hormone production, temperature, mood patterns and the immune system.

Artificial light at night impacts other species, too. Pollinators such as moths, flies and beetles are attracted to lights instead of focusing on feeding, mating or breeding[9]. Bats alter their feeding behaviour; birds, fish and turtles change their migratory routes; and the growth of trees and plants is affected.

CITY LIMITS

The scale of our exposure to artificial lighting is increasing as cities switch sodium street lamps to LEDs. In the United States, 10% of all street lighting has been converted. New York City is changing all 250,000 of its street lights. Milan in Italy was the first city in Europe to do so on such a scale — and the result can be seen from space. By 2015, the city centre's illuminations were brighter and bluer than those of the suburbs.

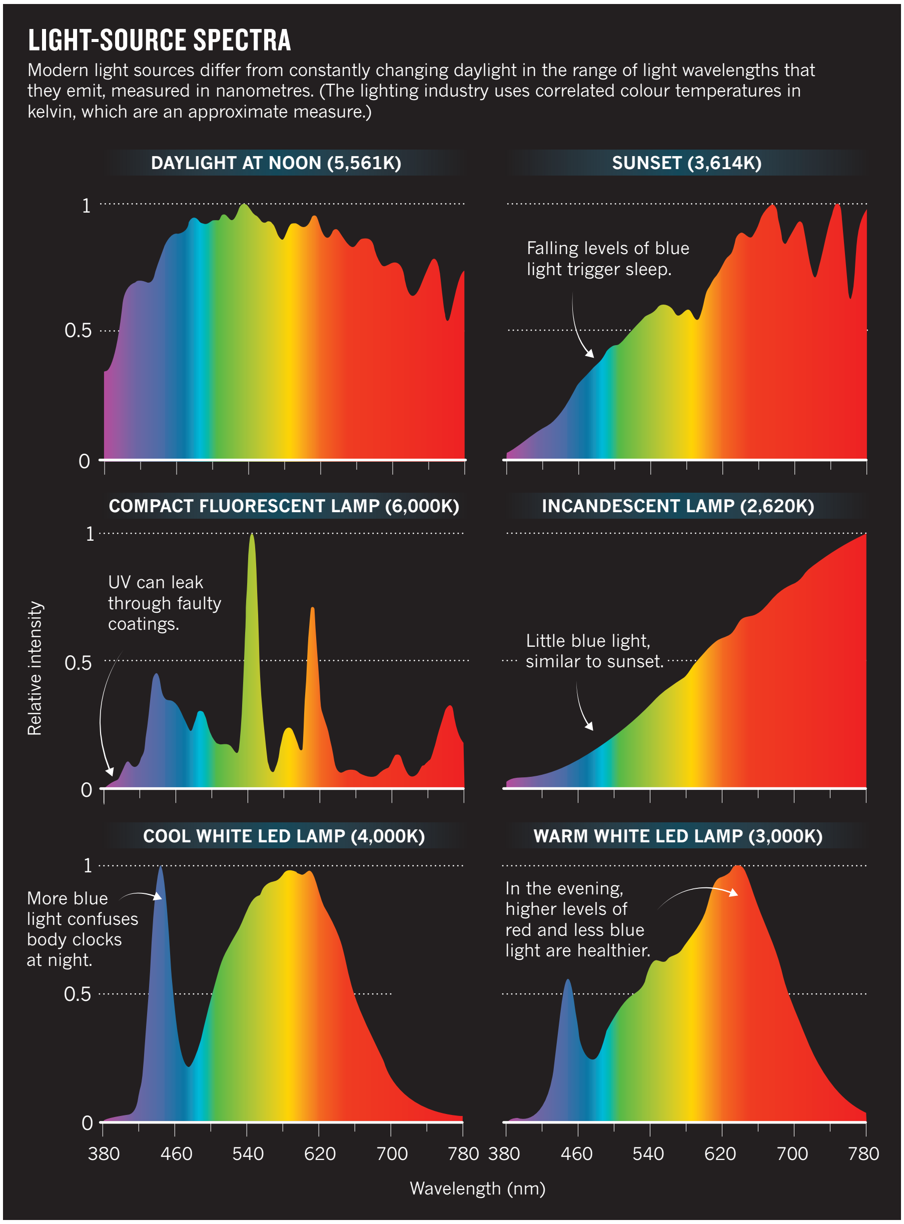
Good lighting design can mitigate some problems. 'Light trespass' into living areas, including bedrooms, can be reduced by designing outdoor luminaires that shine downwards or use shields to block stray rays. Street lights can be dimmed using intelligent control systems and wireless networks of motion sensors. The Van Gogh village in the municipality of Nuenen in the Netherlands, for example, lowers its street lights by 80% when there is no activity and turns them up when a pedestrian, cyclist or car approaches, surrounding them with a safe circle of light as they proceed. Intelligent lighting is expensive to install, but the investment pays back quickly: the Nuenen system reduced energy and maintenance costs by 62%.

New problems requiring regulation are emerging as LEDs become widespread. For example, electromagnetic radiation from wireless lighting controls, outdoor LED signs and digital billboards can interfere with mobile phones, aviation towers and medical equipment such as hearing aids or implantable cardiovascular devices[10].

TIGHTER STANDARDS

Until healthier lighting options become available, the following steps need to be taken to reduce potential negative impacts on the circadian clock. In my opinion, CFLs should be withdrawn from sale because of the scarcity of disposal and recycling protocols. LED sources should be regulated more tightly. Indoors, I recommend using warm white LEDs in the early evening (with colour temperatures below 3,000 kelvin and with as little blue light in the spectrum as possible) and there should be no exposure to light at night, or only to light with a spectrum greater than 600 nm (amber, red colour). Lighting should be indirect, flicker-free and dimmable.

Independent research — beyond the



LIGHT-SOURCE SPECTRA

Modern light sources differ from constantly changing daylight in the range of light wavelengths that they emit, measured in nanometres. (The lighting industry uses correlated colour temperatures in kelvin, which are an approximate measure.)

SOURCE: K. M. ZIELINSKA-DABKOWSKA/ASENSETEK LIGHTING PASSPORT PRO STANDARD SPECTROMETER

lighting industry — is needed into the health and environmental impacts of LED sources, including those with adjustable spectral characteristics, intensity, timing and duration based on the time of the day, evening or night. Emissions outside the visible range must be considered, such as near-infrared radiation (750–950 nm) that is present in daylight and incandescent lamps but not LEDs. Research shows that there needs to be a balance — the use of these light frequencies can repair damaged retinal cells[11] and are necessary. The use of heavy metals in LEDs must be reduced and a process for waste management established. The impacts of control technology in outdoor and indoor spaces must be explored.

Governmental and medical bodies need to draw up stricter regulations and standards for the use of short wavelengths of light at night. In June 2016, the American Medical Association issued a policy statement (Guidance to Reduce Harm from High Intensity Street Lights) to help communities select from the different LED lighting options. Recommendations for light intensity thresholds, timing and duration for indoor and outdoor environments at night are also necessary. It is likewise essential to define the exact spectral characteristics of recommended light sources in nanometres rather than only correlated colour temperatures (CCT) in kelvin. The latter is an approximate measure and cannot accurately describe the light spectrum.

Policymakers should encourage better use of natural light indoors during the day. Artificial light should be used only when there is not enough daylight available, especially in factories, hospitals, nursing homes and offices where people spend a lot of time. Building regulations should reward practices and technologies that harness natural light.

Municipalities should incorporate sustainable night-time illumination polices and guidelines into their urban lighting master plans. Street and security lighting should be directed downwards and shielded. Light levels for walking, cycling and driving should be the minimum acceptable. Passive technologies should be explored. For example, glow-in-the-dark surfaces that absorb energy from the Sun during the day and release it at night could be used on roads and cycle ways (from this low angle, the light would fall on the retinal zone in which blue light has no biological influence). Lights in parks and near forests should be switched off or dimmed late in the evening.

Electromagnetic field emissions from LED outdoor advertisements must be controlled. Digital displays on facades should be no brighter than illuminations on nearby streets, buildings and squares. Installations should be switched off late in the evening to reduce light trespass into residential buildings.

Finally, the public's awareness of lighting issues must be raised. Researchers and lighting practitioners need to communicate the challenges. Healthy lighting design is becoming an important ethical issue that cannot be ignored. An increasing number of communities, such as Monterey in California, are winning lawsuits against municipalities for inappropriate LED city lighting.

For all these reasons, I still use the old incandescent light sources in my home, sleep in complete darkness and spend at least one hour each morning in bright daylight to activate my circadian clock — as do many lighting designers, physicians and chronobiologists. It is imperative that we return to the bright day and dark night cycle that evolution engraved in us. ■ **SEE NEWS FEATURE P.268**

Karolina M. Zielinska-Dabkowska *is a lighting designer, assistant professor at the Faculty of Architecture, Gdansk University of Technology, Poland, and the International Association of Lighting Designers EU Regulatory Affairs Working Group Member.*
e-mail: k.zielinska-dabkowska@pg.edu.pl

1. Gaston, K. J., Visser, M. E., Hölker, F. *Phil. Trans. R. Soc. B* **370,** 20140133 (2015).
2. Naeem, Z. *Int. J. Health Sci. (Qassim)* **4,** 5–6 (2010).
3. James, P. *et al. Environ. Health Perspect.* **125,** 087010 (2017).
4. Rybnikova, N. A., Haim, A. & Portnov, B. A. *Int. J. Obes.* **40,** 815–823 (2016).
5. Cho, J. R., Joo, E. Y., Koo, D. L. & Hong, S. B. *Sleep Med.* **14,** 1422–1425 (2013).
6. Wilkins, A. J., Nimmo-Smith, I., Slater, A. I. & Bedocs, L. *Lighting Res. Technol.* **21,** 11–18 (1989).
7. Behar-Cohen, F. *et al. Progr. Retinal Eye Res.* **30,** 239–257 (2011).
8. Lucas, R. J. *et al. Trends Neurosci.* **37,** 1–9 (2014).
9. Knop, E. *et al. Nature* **548,** 206–209 (2017).
10. de Sousa, M., Klein, G., Korte, T. & Niehaus, M. *Indian Pacing Electrophysiol. J.* **2,** 79–84 (2002).
11. Eells, J. T. *et al. Mitochondrion* **4,** 559–567 (2004).

JOE RAEDLE/GETTY

A DNA-sample library.

GENETICS

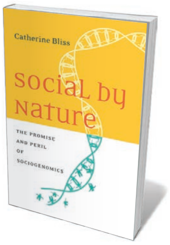
CRISPR's willing executioners

Nathaniel Comfort lauds a sociologist's study of the bias baked into the nature–nurture debate.

In the beginning, there was nature. Then the statistician Francis Galton — Charles Darwin's half-cousin — set nature (heredity) in opposition to nurture, or environment. Galton treated heredity as a family treasure, tucked away in the gametes, shielded from the buffeting environment and passed down the generations. Applying this idea to what he perceived as the degeneration of English manhood, Galton coined a haunting but familiar term: eugenics.

Thus, the nature–nurture binary has been linked with hereditarianism and eugenics from the start. This trio flares up from time to time, for instance in early-twentieth-century eugenics, 1970s sociobiology and the controversial 1994 book on intelligence by Charles Murray and Richard Herrnstein, *The Bell Curve* (Free Press). History doesn't repeat itself, but it winds.

The latest turn of the helix is 'sociogenomics'. This uses genome-wide association studies, high-speed sequencing, gene-editing tools such as CRISPR–Cas9 and baroquely calculated risk scores — often combined with social-science methods — to 'understand' the 'roots' of complex behaviour. In *Social by Nature*, sociologist Catherine Bliss anatomizes the field.



Social by Nature: The Promise and Peril of Sociogenomics
CATHERINE BLISS
Stanford University Press: 2018.

Bliss looks at the science, the professional social structures and the social context of these new developments. She seeks social explanations of why the nature–nurture binary persists in the face of DNA-sequence data that once promised to erase it. Sociogenomics has great biomedical potential, she believes; but the path towards that reward runs along a knife edge, with cliffs of eugenic risk on either side. It is a brilliant book — dense at times, but insightful and filled with illustrative anecdotes and case studies. It's one you should read if you care about what drives academic research, scientific racism or genetic futurism.

Sociogenomics follows many patterns familiar from previous moments of heightened genetic determinism, such as sociobiology, behavioural psychology or the debate ignited by *The Bell Curve*. But Bliss argues that, this time, it's different. She suggests that genetic methods have never promised so much, while delivering so little. As a historian, I see more consistency in the promises of human genetics over time; nevertheless, Bliss's findings are striking.

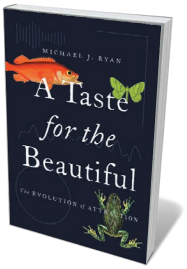
She notes, for example, a special issue of the journal *Biodemography and Social Biology* from 2014 (see go.nature.com/2qnovjh) concerning risk scores. (These are estimates of how much a one-letter change in the DNA code, or SNP, contributes to a particular disease.) In the issue, risk scores of between 0% and 3% were taken as encouraging signs for future research. Bliss found that when risk scores failed to meet standards of statistical significance, some researchers — rather than investigate environmental influences — doggedly bumped up the genetic significance using statistical tricks such as pooling techniques and meta-analyses. And yet the polygenic risk scores so generated still accounted for a mere 0.2% of all variation in a trait. "In other words," Bliss writes, "a polygenic risk score of nearly 0 percent is justification for further analysis of the genetic determinism of the traits". If all you have is a sequencer, everything looks like an SNP.

"Sociogenomics has great biomedical potential, but the path towards that reward runs along a knife edge."

What the historian Andrew Hogan has called the "genomic gaze" isn't the fault of individual bad-guy researchers: it's structural. Bliss is careful to acknowledge the good, even noble intentions of many of the scientists she spoke to (as a sociologist, she keeps the names of her 'informants' confidential). But she finds that the funding and publicity mechanisms integral to biology drive it towards genes-first explanations. The stakes are high: finding an SNP associated with a risk increase from 0.01% to 0.03% (a threefold rise) for a disease such as breast cancer could make a career. "While researchers do not intend to lift the focus off of the environment," Bliss writes, "they are forced to recast social phenomena as 'evolutionary phenotypes' so that they can make scientific claims" that sound relevant to biomedical funders.

▶

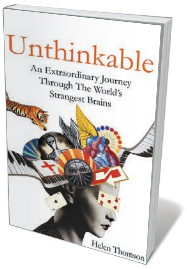
Books in brief



A Taste for the Beautiful: The Evolution of Attraction

Michael J. Ryan PRINCETON UNIVERSITY PRESS *(2018)*

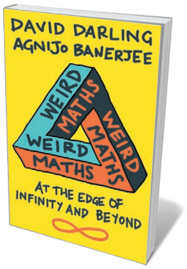
In terms of sexual selection, the iridescent bling of a peacock's tail is just another lure in an array of animal arias, odours and ornaments. And as Michael Ryan argues in this lucid study, such beauties reside "in the brain of the beholder". Kicking off with his research on the tiny Central American túngara frog (*Engystomops pustulosus*), the males of which emit a complex call, Ryan examines sexual beauty in all its sensory forms, as well as fickleness, hidden preferences and experiments with quail that could shed light on the predilection for pornography.



Unthinkable

Helen Thomson JOHN MURRAY *(2018)*

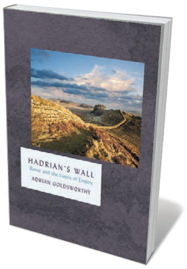
Botched surgery, accidents, mutations and disease: as Helen Thomson reminds us in this exploration of rare neurological conditions, trauma has told us much about the brain. She neatly integrates sensitive interviews with patients into current research on their conditions and historical case studies. We meet, for instance, Sharon, who cannot generate mental maps and feels permanently 'lost'; and Graham, who believed he was dead (Cotard's syndrome) for three years. The result is a stirring scientific journey, a celebration of human diversity and a call to rethink the 'unthinkable'.



Weird Maths

David Darling and Agnijo Banerjee ONEWORLD *(2018)*

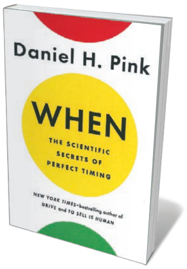
This frolic on the wilder shores of mathematics, by astronomer David Darling and maths prodigy Agnijo Banerjee, aims to bolt the way-out to the day-to-day. It succeeds. After playing with the question of whether the cosmos is innately mathematical, or just looks that way, Darling and Banerjee plunge into the deep. Here, they surf the big waves: invigorating concepts such as how to see in four dimensions, the inner structure of the Mandelbrot set of fractals, the musical scales of alien cultures, Georg Cantor's work on hierarchies of infinity and the uncomputably huge Rayo's number.



Hadrian's Wall

Adrian Goldsworthy HEAD OF ZEUS *(2018)*

Stretching just over 100 kilometres coast to coast across the north of England, Hadrian's wall is a pipsqueak compared to the Great Wall of China. Yet the barrier, begun in AD 122, is a stunning testament to Roman engineering in a far-flung corner of the empire. As historian Adrian Goldsworthy explains in this succinct study, its real purpose (protection from Picts, or display of power?) remains enigmatic, but much else is known. He follows the emperors who put their stamp on 'Britannia', and explores the wall and its garrisons up to the fifth century, when Germanic tribes fatally disrupted Roman rule.



When

Daniel H. Pink RIVERHEAD *(2018)*

When is the best time to start a relationship, change career or eat dinner? Daniel Pink analysed 700 studies in anthropology, endocrinology, social psychology and beyond to probe the science of timing. He unpicks compelling patterns: why medical malpractice and harsher judicial rulings cluster in the afternoon; how we pay too much attention to endings; which circumstances favour synchronization in teams. And he includes handy 'time-hacking' advice on how to put the insights divulged into practice. **Barbara Kiser**

ROBERT FRANKLIN/SOUTH BEND TRIBUNE VIA AP

Students at the University of Notre Dame in Indiana protest outside an event featuring the author of a controversial book on intelligence.

This tendency has social implications. 'Just-so stories' abound, reinforcing toxic stereotypes. For example, Bliss cites peer-reviewed work speculating that violence might get men more sex. And prevention can grade into genetic surveillance: after the 2012 mass shooting at Sandy Hook Elementary School in Newtown, Connecticut, the state asked a geneticist to examine gunman Adam Lanza's genome for markers that might have predisposed him to violence.

Bliss handles sensitive categories such as race, gender and sexuality with subtlety, examining the interplay of peer-reviewed articles and their media coverage. For example, she notes that most social-genomics papers "make rote references to racial differences without defining what they mean". She observes that mass-culture gender norms, by contrast, inflect peer-reviewed articles, demonstrating that culture shapes science as well as the reverse.

Some of Bliss's informants even contemplate the creation of DNA-based social strata. "You know," one reports a colleague saying, "it'll be great when we can have the janitors just be janitors." Shades of Aldous Huxley's *Brave New World*: I'm so glad I'm a Beta.

"Genetic determinism isn't just spread over genomics like poisoned icing. It's baked into how we do research."

Genetic determinism, then, isn't just spread over genomics like poisoned icing. It's baked into how we fund, conduct and disseminate research. Unlike the optimists who claim that individualism and the free market immunize us against eugenic evils, Bliss sees both as rife with eugenic risk. The medical marketplace helps to reify the idea that your genome is your true identity. It lends scientific authority to efforts to find 'objective' answers to impossible, hopelessly social questions about, say, IQ. Direct-to-consumer advertisements often target children or parents. The Children's Palace in Chonqing, China, for instance, hosts a "genetics summer camp" for children aged 3–12 that claims to identify and then to develop 'traits' such as sporting and musical ability.

I'm less convinced than Bliss that this genocentrism is new to the genomic age. I readily concede that genomics gives new power to hereditarian explanations of human behaviour, and that our culture is newly conducive to 'gene-for' research. But much of what she describes sounds to me like determinism in a new context.

What Bliss does brilliantly is analyse the mechanisms by which genetic determinism is an outcome of the research endeavour itself. Her most searing conclusion is that scientists and journalists can understand that nature and nurture are not zero-sum, can even strive to strike essentialist language from their work, and yet can still serve the god of genetic determinism. Driven by capital, individualism and the lure of interdisciplinarity, we may be opposed to the ideology and yet willingly participate in its prosecution. In historical context, that is a haunting thought. ■

Nathaniel Comfort *is professor of the history of medicine at Johns Hopkins University in Baltimore, Maryland, and is the author, most recently, of* The Science of Human Perfection. *He is working on a biography of DNA.*
e-mail: nccomfort@gmail.com

Correspondence

Research kudos does not need a price tag

We are concerned that the focus on generating grant income to fill university coffers penalizes science that is good value for money. As research money dwindles, the winning of funds seems to be emerging as the way to judge performance. This criterion is used by the UK universities' Research Excellence Framework, for example, and to assess researchers for hiring or promotion.

Huge strides are being made in our field of whole-organism biology owing to large collaborative grants that help to pay for salaries and equipment. However, important oases of the biological sciences are relatively unaffected by the benefits of big grants. Examples include biological modelling, ecological projects in the developing world and meta-analyses based on literature mining.

At a time when economic efficiency is paramount in publicly funded areas such as health and education, an undue emphasis on generating money could drive scientists to compete for limited public funds simply for career purposes. In our view, funding success must not become a disproportionate factor in gauging scientific achievement.

Tim Caro *University of California, Davis, USA.*
Sasha R. X. Dall *University of Exeter, Penryn, Cornwall, UK.*
tmcaro@ucdavis.edu

Don't misrepresent bats' link with SARS

We find your report on bats and severe acute respiratory syndrome (SARS) sensationalist and misleading (*Nature* **552,** 15–16, 2017). The important work it discusses does not claim to pinpoint conclusively the source of the SARS outbreak (B. Hu *et al. PLoS Pathog.* **13,** e1006698; 2017), as implied by your "smoking gun" metaphor. The rapid rate of evolution of RNA viruses means that SARS could have arisen in one of many areas. Thus, your inference that the strain "could easily" have originated in this bat population is, in our view, unjustified.

Inflammatory statements about bats and disease have led to culling and roost destruction, compromising conservation efforts (K. J. Olival *EcoHealth* **13,** 6–8; 2016). Accurate reporting of information on SARS, Middle East respiratory syndrome, Ebola and other emerging diseases is crucial for controlling outbreaks and for preventing unnecessary deaths of wild animals.

Viral spillover occurs when humans and domestic animals come into direct contact with wild animals and their pathogens. Public education, comprehensive surveillance and considered interventions can all help to protect public health. The closure of markets selling live birds has already reduced the activity of avian influenza viruses, and could likewise curtail the spillover of mammalian viruses.

Paul A. Racey* *University of Exeter, Penryn, Cornwall, UK.*
p.a.racey@exeter.ac.uk
**On behalf of 5 correspondents (see go.nature.com/2qzjxt4 for full list).*

Train robots to self-certify as safe

Robots can operate autonomously in extreme environments that might be hazardous for humans. For example, they can inspect oil and gas equipment, monitor offshore wind turbines, survey subsea power networks and maintain nuclear reactors. We suggest that these robots should be required to self-certify that they can operate safely under such circumstances.

Robots can learn to adapt the way they perform tasks in changing and unexpected environments. However, if a robotic system learns a flawed model of the environment or a risky behaviour, it could undermine its own operation and the integrity of the asset that it is inspecting or repairing — with potentially catastrophic consequences. To protect against this, the robot should self-certify its correct operation by collecting data as it executes its task. It would then check the data against its mission plan, with minimal input from human operators (see D. M. Lane *et al. IFAC Proc. Vol.* **45,** 268–273; 2012).

For autonomous systems to be trusted, developments in robotics and artificial intelligence need to be accompanied by advances in certification techniques. Regulators such as Lloyd's Register have certification standards for industrial equipment and are beginning to explore the challenges of certifying self-learning systems (see go.nature.com/2cxxjcx). And several teams at Research Councils UK, including the Offshore Robotics for Certification of Assets hub (https://orcahub.org), are investigating this crucial area.

Valentin Robu, David Flynn, David Lane *Heriot-Watt University, Edinburgh, UK.*
v.robu@hw.ac.uk

Statistics: a social and cultural issue

Too many practitioners who discuss the misuse of statistics in science propose technical remedies to a problem that is essentially social, cultural and ethical (see J. Leek *et al. Nature* **551,** 557–559; 2017). In our view, technical fixes are doomed.

As Steven Goodman writes in the article, there is nothing technically wrong with *P* values. But even when they are correct and appropriate, they can be misunderstood, misrepresented and misused — often in the haste to serve publication and career. *P* values should instead serve as a check on the quality of evidence.

The great paradox of science is that passionate practitioners must carefully produce dispassionate facts (J. Ravetz *Scientific Knowledge and its Social Problems* Oxford Univ. Press; 1971). Meticulous technical and normative judgement, as well as morals and morale, are necessary to navigate the forking paths of the statistical garden.

Unless peer review and rewards in academia change to encourage such virtues, the present crisis will remain intractable (see also A. Saltelli and S. Funtowicz *Futures* **91,** 5–11; 2017).

Andrea Saltelli *University of Bergen, Norway.*
Philip Stark *University of California, Berkeley, USA.*
andrea.saltelli@uib.no

Statistics: deploy with integrity

Discussions to strengthen the quality of statistical analyses are a welcome demonstration of scientists' willingness to confront uncomfortable knowledge (J. Leek *et al. Nature* **551,** 557–559; 2017). Just as science in general is not a truth machine, statistics is not a device for automatically bootstrapping certainty out of data sets.

All users of statistical techniques, as well as those in other mathematical fields such as modelling and algorithms, need an effective societal commitment to the maintenance of quality and integrity in their work. If imposed alone, technical or administrative solutions will only breed manipulation and evasion.

There may be methodological issues as well. For example, we are only now discovering that the universally accepted standard tests, notably significance and *P* values, are simplistic and misleading. It might be that improved tests, such as those involving power calculations, are just too sophisticated for otherwise competent researchers. If so, then the conduct of empirical science will need substantial modification.

Jerome Ravetz *University of Oxford, UK.*
jerome.ravetz@gmail.com

Ben Barres

(1954–2017)

Neurobiologist who advocated for gender equality in science.

Ben Barres (born Barbara Barres) was a passionate researcher of the role of glia, the most numerous type of brain cell, in development and disease. He was also an ardent campaigner for equal opportunity in science. He died of cancer aged 63, on 27 December 2017.

As Barbara and as Ben (he transitioned genders in 1997), Barres made numerous landmark discoveries. These include the identification of glial-derived factors that promote the formation and elimination of synapses, and the characterization of signals that induce the formation of myelin, the lipid sheathing on neurons.

Barres devoted much of his last decade to publicly describing the challenges he had faced as a woman in science, and offering ways to correct a system that he viewed as fundamentally biased against the advancement of women and minorities. He also called for mentors to be held more accountable for the training and success of their graduate students and postdocs.

Barres took tremendous pleasure from working on important but neglected problems. "Ninety-nine per cent of neuroscientists work on 1% of the interesting questions," he said, "It is so much more exciting to work on the untouched mysteries!" His findings and vocal presence at meetings were largely responsible for the acceptance that glial cells contribute to brain development, function and disease.

Barres was raised in West Orange, New Jersey and loved mathematics and science from an early age. He never felt comfortable being treated as a girl. At school, Barres repeatedly requested, but was denied, access to courses in science and engineering. A summer science programme with no gender restrictions at Columbia University in New York City finally provided access to these subjects, and led him to pursue a bachelor of science degree in biology at the Massachusetts Institute of Technology in Cambridge.

In 1979, Barres completed a medical degree at Dartmouth College in Hanover, New Hampshire, and then a neurology residency at Weill Cornell Medicine in New York City. He was intrigued that so many of the diseases that impair brain and nervous-system function involve glial cells, yet so little was known at the time about their biology.

STANFORD SCHOOL OF MEDICINE

Barres left medicine to do a doctorate in neurobiology at Harvard Medical School in Boston, Massachusetts, on the function and distribution of cation channels in glial cells. During a postdoc at University College London, Barres discovered that developing neurons provide signals to the myelinating glial cells — the oligodendrocytes — to insulate neuronal axons.

Barres started his own lab in 1993, in the neurobiology department at Stanford University School of Medicine in California. He mentored dozens of students and postdocs. Barres' lab meetings were legendarily intense. They often lasted three hours or more owing to the large number of people in attendance, the vast range of topics covered and the open, somewhat unstructured discourse that Barres encouraged.

Barres insisted that people in his lab tackle important scientific problems, voice their views and ask questions at conferences. He allowed his trainees immense freedom to collaborate with others, to work whatever hours they wanted and to attend any meeting they cared to — as long as they asked questions. Many went on to faculty positions in the United States, Europe or Asia.

The Barres lab made many discoveries about how synapses form in the developing brain. It probed the roles of different types of glial cell — astrocytes and microglia — in synapse elimination, for example. Barres also made significant contributions to the study of signals that influence the survival of damaged neurons, optic-nerve and spinal-cord regeneration, and the assembly and maintenance of the barrier that prevents specific molecules in the blood entering the brain. In 2013, Barres was elected to the US National Academy of Sciences — the first openly transgendered member.

Ben had an almost superhuman work ethic. Working 18–20 hours per day didn't feel difficult, he told me, because, "science is fun ... almost like a playful addiction". In his later years, Ben started cycling in the hills around Stanford. He also began roasting his own coffee beans — giving bags to lab members in exchange for constructive feedback. He delighted in all things Harry Potter. Trips to the latest film were among the few mandatory requirements of Barres-lab membership.

In April 2016, Ben was diagnosed with advanced pancreatic cancer. Amid repeated treatments, he continued to work every day, write grant applications and manuscripts, and keep up his advocacy. He never stopped mentoring his students and postdocs, and toiled feverishly to update and archive their letters of support in anticipation of their future career developments after his death.

Ben confided that he'd never had much interest in romantic relationships or having children. He told me: "I've always considered my colleagues as my family, and my students and my postdocs as my children." Seeing them flourish and succeed was one of his greatest sources of joy.

Ben's colleagues and protégés adored him and considered the Barres lab a family of sorts, too. As far back as I can recall, the hallway doors of Ben's lab were adorned with drawings and photos of the various lab members — past and present, their children and their pets and, of course, glial cells. We were and remain bonded by our affection and appreciation for Ben, his dedication to mentoring us, his quirks, and his unrelenting spirit.

Barres was remarkably brave and reflective about his illness and the life he'd led. As he put it: "I lived life on my terms: I wanted to switch genders, and I did. I wanted to be a scientist, and I was. I wanted to study glia, and I did that too. I stood up for what I believed in and I like to think I made an impact, or at least opened the door for the impact to occur. I have zero regrets and I'm ready to die. I've truly had a great life." ■

Andrew D. Huberman *is a professor of neurobiology at Stanford University School of Medicine, California, and a former postdoc in the Barres lab. He was a close friend of Ben's.*
e-mail: adh1@stanford.edu

COGNITIVE NEUROSCIENCE

Mice learn to avoid the rat race

Mice can learn to overcome their naturally aggressive approach to conflict resolution, instead adopting a cooperative strategy. This discovery provides a simple animal model in which to investigate a complex social behaviour.

SCOTT M. RENNIE & MICHAEL L. PLATT

Social interactions are often complicated by conflicts of interest. Humans and other animals adopt diverse strategies to resolve such disputes. Stronger individuals can often secure their interests at the expense of weaker individuals, but this strategy can be costly if it requires aggression. Strategies that are more cooperative and egalitarian can also develop among kin[1] or individuals who reciprocate in repeated interactions[2]. Theoretical and experimental studies suggest that cooperation depends on cognitive control processes that override the impulse to acquire tangible rewards[3]. This theory now finds support from Choe *et al.*[4], writing in *Nature Communications*. The authors demonstrate that pairs of mice can learn to coordinate their behaviour to achieve an egalitarian distribution of rewards — but only when rewards are delivered directly to the brain, rather than through food.

Mice are flexible in their social behaviour. At low population densities, they establish and aggressively defend territories, whereas at higher population densities, they develop strict hierarchies in which a single male dominates several subordinates[5]. Neither of these strategies whiffs of cooperation. For male mice, as for many other animals, size, aggressiveness and persistence strongly determine social rank. Mice decide whether to compete by comparing potential costs and benefits on the basis of perceived asymmetries in these qualities. These computations rely on a neural circuit that connects two brain regions — the mediodorsal thalamus and the dorsomedial prefrontal cortex[6].

Choe *et al.* set out to investigate whether mice have the capacity to override their natural tendencies towards dominance-based conflict resolution. To do this, they developed a clever coordination task. They trained mice to enter a central start zone in a three-chambered box, and then to follow a visual cue to either the left or right chamber of the box to receive a reward. Next, they paired trained animals to take the trial together. When both mice occupied the start zone, a trial was initiated (Fig. 1a). The first mouse to enter the correct chamber received a reward of either food pellets or wireless brain stimulation (WBS) of the medial forebrain bundle — a region that, when stimulated, can override all other rewards, including food, water and sex[7] (Fig. 1b). In the WBS trials, the reward was terminated if the second mouse entered the chamber (Fig. 1c), although this was not possible in the food trial.

As expected, when mice were rewarded with food pellets, dominant ones coerced their subordinate partners into the start zone to enable the trial to begin, and then monopolized the rewards. By contrast, Choe and colleagues found that most animals that were rewarded with WBS developed and maintained a simple alternate-side-allocation rule: each mouse in a pair monopolized only one reward chamber and avoided the other (Fig. 1d). As a result, one mouse gained rewards in trials when the left-hand chamber was the reward chamber, and the other gained rewards when the right-hand chamber was the reward chamber. By following this rule, mice increased both the total amount of reward received and the equality with which that reward was divided.

Remarkably, WBS seemed to override the hierarchical, despotic behaviour that

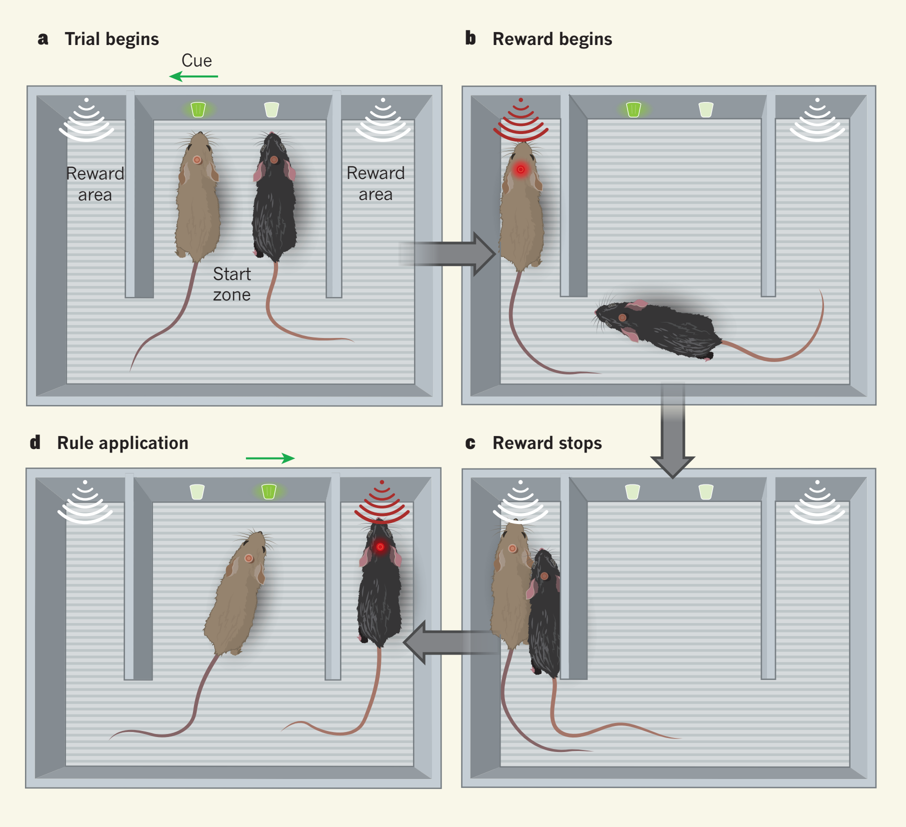


Figure 1 | Mice learn a rule to resolve conflict. **a**, Choe *et al.*[4] taught mice that entering a start zone in a three-chambered box would trigger a trial, in which a visual (light) cue would indicate where to go to receive a reward — into either the left- or right-hand reward area. Once trained, mice performed the task in pairs. Both animals had to occupy the start zone to initiate a trial. **b**, The first animal to enter the cued reward zone received a reward of wireless brain stimulation of the medial forebrain bundle (red circle). **c**, If the second mouse entered the reward zone, the reward was terminated. **d**, The authors showed that the pairs learnt to observe a side-allocation rule, in which one followed the cue to rewards in the left-hand chamber, the other to those in the right. This increased overall reward and equity between the animals.

developed over food rewards. Rule-following mice in WBS trials displayed very little aggression, and the limited aggression observed had minimal impact on choice behaviour. Asymmetries in the sizes of the paired animals, which are a key determinant of social status, also had no effect on WBS-induced cooperation. Even when the authors reshuffled the mice into new pairs in which both animals had the same side preference (both monopolizing the left chamber in their previous trials, for instance), the animals rapidly re-established the alternate-side-allocation rule — thus demonstrating remarkable flexibility.

Choe and colleagues' experiments indicate that certain factors can put natural limitations on cooperation. These include food deprivation[8] and the presence of a powerful appetitive stimulus, the food pellet, which was clearly visible in the food-reward trials, and was presumably aromatic, too. By contrast, although WBS was associated with a light cue, it was otherwise not obvious to the unrewarded animal. These findings resonate with previous studies showing that the physical presence of tangible rewards impairs delayed gratification in blue jays[9], complex rule-following by monkeys[10] and chimps[11], and cooperation in humans[12].

The current study raises several questions. First, is social coordination by rule-following supported by the same neural circuit between the mediodorsal thalamus and dorsomedial prefrontal cortex that underlies status-based conflict resolution? If not, perhaps WBS overrides this circuit by triggering different circuits that stamp in a more 'cognitive' strategy.

Second, what role do internal states, such as hunger, have in strategy selection? The mice in Choe and colleagues' WBS trials were not food-deprived, and it would be interesting to determine how hunger would affect their behaviour.

And third, to what extent is rule-based coordination social at all? Determining to what extent this coordination depends on physical similarity between partners, transmission of social signals, or the implementation of a similar computational routine could provide clues to this question. If WBS could elicit the same type of coordination between a mouse and a robot, for example, this would demonstrate that the behaviour observed in the current study does not involve any sort of attribution of agency or strategic thinking, and instead arises from pure associative learning. Could WBS drive cooperation between the cartoon characters Tom the cat and Jerry the mouse, or might it just stop them from fighting?

Choe *et al.* have provided a compelling demonstration of a transition from aggressive to more-egalitarian interactions, at a time when examples of cooperation between animals in the laboratory are controversial and rare[8]. Crucially, they have done so in mice, an animal model that will allow the whole range of powerful techniques in the neuroscience toolbox, from behaviour tracking to molecular-genetic tools such as optogenetics to electrophysiology, to be brought to bear on the these tricky but important social questions. ■

Scott M. Rennie *is in the Champalimaud Neuroscience Programme, Champalimaud Centre for the Unknown, 1400-038 Lisbon, Portugal.* **Michael. L. Platt** *is in the Departments of Neuroscience, of Psychology and of Marketing, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.*
e-mails: scott.rennie@neuro.fchampalimaud.org; mplatt@pennmedicine.upenn.edu

1. Hamilton, W. D. *Am. Nat.* **97,** 354–356 (1963).
2. Trivers, R. L. *Q. Rev. Biol.* **46,** 35–57 (1971).
3. Stevens, J. R. & Hauser, M. D. *Trends Cogn. Sci.* **8,** 60–65 (2004).
4. Choe, I.-H. *et al. Nature Commun.* **8,** 1176 (2017).
5. Wang, F., Kessels, H. W. & Hu, H. *Trends Neurosci.* **37,** 674–682 (2014).
6. Zhou, T. *et al. Science* **357,** 162–168 (2017).
7. Olds, J. & Milner, P. *J. Comp. Physiol. Psychol.* **47,** 419–427 (1954).
8. Viana, D. S., Gordo, I., Sucena, E. & Moita, M. A. P. *PLoS ONE* **5,** e8483 (2010).
9. Stephens, D. W., McLinn, C. M. & Stevens, J. R. *Science* **298,** 2216–2218 (2002).
10. Silberberg, A. & Fujita, K. *J. Exp. Anal. Behav.* **66,** 143–147 (1996).
11. Boysen, S. T. & Berntson, G. G. *J. Exp. Psychol. Anim. Behav. Processes* **21,** 82–86 (1995).
12. Nishi, A., Shirado, H., Rand, D. G. & Christakis, N. A. *Nature* **526,** 426–429 (2015).

This article was published online on 21 December 2017.

MICROBIOLOGY

Pathogens boosted by food additive

Epidemic strains of the bacterium *Clostridium difficile* have now been found to grow on unusually low levels of the food additive trehalose, providing a possible explanation for *C. difficile* outbreaks since 2001. SEE ARTICLE P.291

JIMMY D. BALLARD

Between 2001 and 2006, epidemic strains of the bacterium *Clostridium difficile*, which can inhabit the bowel and cause dangerous diarrhoea, unexpectedly emerged in the United States, Canada and several European countries[1,2]. Most of these strains originated from a single lineage of *C. difficile* known as ribotype 027 (RT027; ref. 2), which has now spread around the world[3]. Of particular concern has been the correlation between RT027 and a dramatic increase in deaths related to *C. difficile*[4]. The mystery of why this ribotype and a second one, RT078, became so prevalent apparently out of thin air has remained largely unsolved[5]. On page 291, Collins *et al.*[6] raise the possibility that the seemingly harmless addition of a sugar called trehalose to the food supply contributed to this disease epidemic.

Collins and colleagues first explored how RT027 and RT078 grow, by comparing carbon-source preferences between strains of *C. difficile*. They noted a peculiar property of these two lineages — they can use low concentrations of trehalose as a sole source of carbon. Next, the authors analysed the genomes of RT027 and RT078, and discovered that each encodes unusual sequences that might explain their ability to grow in low levels of trehalose.

The researchers showed that RT027 carries a single-nucleotide genetic variant that changes an amino-acid residue in the protein TreR from leucine to isoleucine. TreR is a transcriptional repressor that is inhibited by trehalose. When active, TreR prevents expression of the gene *treA*, which encodes a phosphotrehalase enzyme involved in metabolizing trehalose into glucose and glucose derivatives. Thus, trehalose is metabolized only when its levels are high enough to inhibit TreR. Collins *et al.* propose that the mutation in RT027 changes TreR's affinity for trehalose and allows it to be repressed by substantially lower levels of the sugar than normal. This frees the TreA protein to metabolize trehalose and allows RT027 to grow on low levels of the sugar (Fig. 1).

By contrast, RT078 has adapted to grow on low amounts of trehalose by acquiring four genes involved in trehalose uptake and metabolism. The genes encode second copies of TreR and TreA, a trehalose transporter protein dubbed PtsT that helps cells take up the sugar, and another enzyme, TreX, involved in trehalose metabolism. Unexpectedly, RT078 does not share the genetic alteration in TreR that is found in RT027. As Collins and colleagues point out, it therefore seems that two epidemic strains of *C. difficile* have optimized trehalose metabolism in unrelated ways.

The investigators next provided evidence that trehalose metabolism directly relates to enhanced virulence of RT027 *in vivo*. First, they showed that deleting *treA* in RT027 and thereby preventing trehalose metabolism

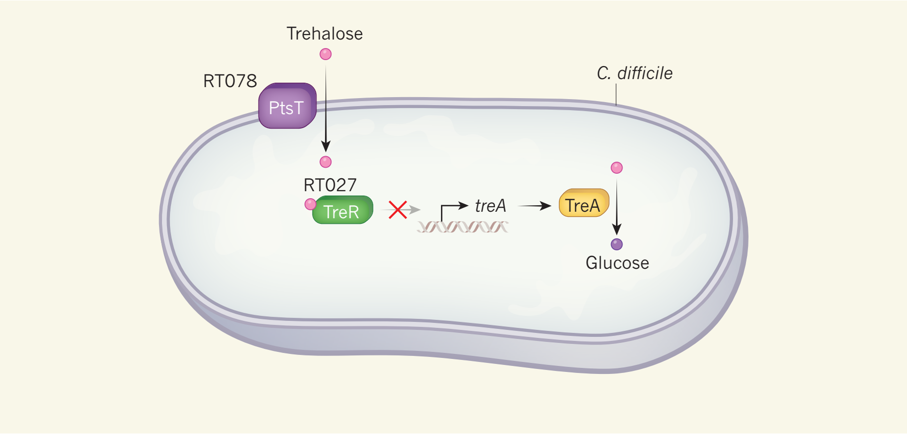


Figure 1 | Increased virulence of the bacterium *Clostridium difficile*. Two lineages of *C. difficile*, RT027 and RT078, have become widespread since the early 2000s. Collins *et al.*[6] have demonstrated that different mutations have arisen in each strain to improve the microbes' ability to grow on low concentrations of the sugar trehalose, which has been added to foods since 2001. RT078 has acquired four genes, including one that encodes the protein PtsT, which transports trehalose into *C. difficile* cells. In RT027, mutation of the protein TreR increases the protein's affinity for trehalose, which in turn inhibits TreR's ability to bind to DNA and repress transcription of the gene *treA*. TreA protein, expressed when TreR is repressed, metabolizes trehalose to glucose and derivatives, enabling cell growth at low trehalose concentrations.

markedly reduced the virulence of this strain in mice. Second, adding trehalose to the diet of mice infected with RT027 increased the animals' risk of death. However, the bacterial load of RT027 was not higher in mice fed trehalose than in those on a trehalose-free diet, indicating that increased risk of death is not simply due to the presence of more bacteria. Rather, the authors found that improved trehalose metabolism enables RT027 to produce higher levels of a *C. difficile* toxin.

Turning to RT078, Collins *et al.* demonstrated that just one of the four acquired proteins, the trehalose transporter PtsT, was responsible for the strain's increased ability to grow on low levels of trehalose (Fig. 1). The authors showed that PtsT confers a competitive growth advantage over other lineages in the presence of trehalose.

Finally, Collins and colleagues investigated the relevance of their observations in humans. Experimental infection would be difficult in people, so the researchers instead collected fluid from the small intestine of three participants on a normal diet. The fluid contained levels of trehalose sufficient to promote expression of *treA* in RT027 but not in other strains, supporting the potential for human relevance.

RT027 was first isolated in 1985, from a person infected with *C. difficile*. But this ribotype was not associated with hospital outbreaks, increased death rates or epidemics until the early 2000s. Similarly, RT078 lineages isolated before the *C. difficile* epidemics carry the genetic information for enhanced trehalose metabolism, but were of little consequence to the epidemiology of this disease. Why did these ribotypes suddenly emerge at epidemic levels only 15 years ago?

Collins and colleagues propose a surprising answer. Before 1995, high production costs made trehalose untenable as a food additive. But manufacturing innovations[7] reduced the cost of trehalose production more than 100-fold[8], and the US Food and Drug Administration and European agencies approved the sugar as a safe food additive in 2000 and 2001, respectively (see go.nature.com/2yewlwk; go.nature.com/2jyltr3). Trehalose is now added to a variety of food products, including pasta, ice cream and minced beef. The authors provide a timeline (see Fig. 6 of the paper) to illustrate how supplementing the food supply with trehalose preceded the *C. difficile* outbreaks caused by RT027 and RT078. They therefore suggest that the addition of trehalose to the food supply might have increased the sugar in the human bowel to levels high enough to enable growth of these ribotypes.

The study's findings raise several avenues for future research. For instance, the connection between trehalose metabolism and toxin production, and how this is linked to increased death rates in people infected with RT027, will require further analysis. Whether trehalose in the human colon, where disease occurs, reaches high enough levels to affect RT027 and RT078 virulence is also unknown. The authors tested fluid from the small intestine, thus bypassing the colon, where the complex complement of gut microbes might break down trehalose.

Despite these concerns, the correlative findings of Collins and colleagues' study are compelling. It is impossible to know all the details of events surrounding the recent *C. difficile* epidemics, but the circumstantial and experimental evidence points to trehalose as an unexpected culprit. ■

Jimmy D. Ballard *is in the Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma 73190, USA.*
e-mail: jimmy-ballard@ouhsc.edu

1. Bartlett, J. G. *Ann. Intern. Med.* **145,** 758–764 (2006).
2. McDonald, L. C. *et al. N. Engl. J. Med.* **353,** 2433–2441 (2005).
3. He, M. *et al. Nature Genet.* **45,** 109–113 (2013).
4. Hunt, J. J. & Ballard, J. D. *Microbiol. Mol. Biol. Rev.* **77,** 567–581 (2013).
5. Kuijper, E. J., van Dissel, J. T. & Wilcox, M. H. *Curr. Opin. Infect. Dis.* **20,** 376–383 (2007).
6. Collins, J. *et al. Nature* **553,** 291–294 (2018).
7. Maruta, K. *et al. Biosci. Biotechnol. Biochem.* **59,** 1829–1834 (1995).
8. Higashiyama, T. *Pure Appl. Chem.* **74,** 1263–1269 (2002).

This article was published online on 3 January 2018.

CHEMICAL BIOLOGY

Strategy for making safer opioids bolstered

Compounds have been made that activate only the G-protein signalling pathway when bound to the μ-opioid receptor — the target of opioid pain relievers. These compounds lack one of the main side effects of currently used opioids.

SUSRUTA MAJUMDAR & LAKSHMI A. DEVI

Effective pain management is one of the greatest challenges of modern medicine. Opioids such as morphine and fentanyl are the preferred clinical treatments for moderate to severe pain because of their strong analgesic (pain-relieving) effects. But the ongoing epidemic of deaths from respiratory depression induced by opioid overdoses highlights the need for safer analgesics. Writing in *Cell*, Schmid *et al.*[1] report a series of compounds that provides a much-needed proof of principle of a strategy for making safer opioids.

It is estimated that more than 100,000 adults suffer from chronic pain in the United States alone, and that this costs up to US$635 billion per year in medical treatment and lost workforce productivity[2]. The most commonly used drugs for pain management include opioids, non-steroidal anti-inflammatory drugs (NSAIDs) and paracetamol (acetaminophen). However, these treatments can have numerous side effects. For example, NSAIDs can cause cardiovascular complications, gastrointestinal bleeding and renal disease, and acetaminophen is toxic to the liver.

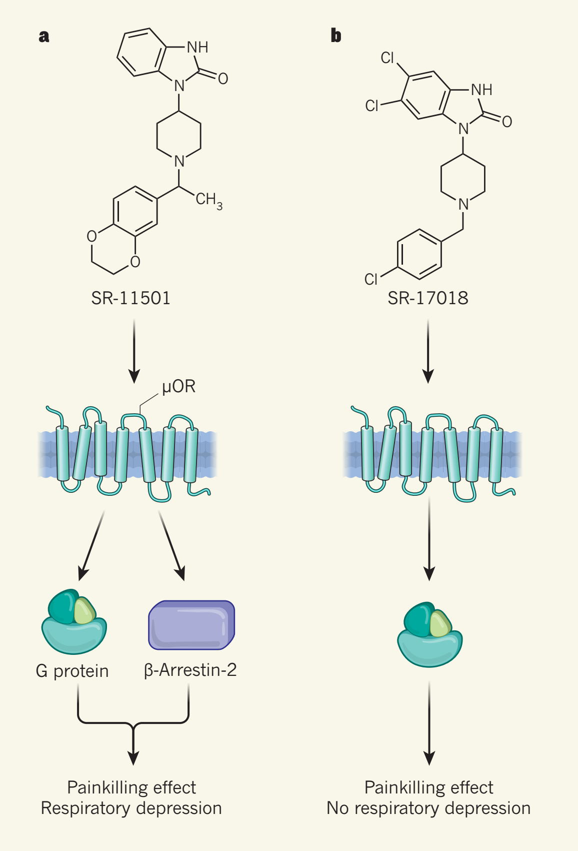


Figure 1 | Signalling preferences correlate with the physiological effects of opioids. Schmid *et al.*[1] prepared compounds that bind to and activate the μ-opioid receptor (μOR) — the biological target of opioid painkillers. **a**, μOR activation by some compounds, such as SR-11501, triggers signalling through two pathways: the G-protein pathway and the β-arrestin-2 pathway. Such compounds are pain relievers in mice, but cause respiratory depression as a side effect. **b**, By contrast, compounds such as SR-17018, which trigger only G-protein-mediated signalling, retain their pain-relieving activity but do not cause respiratory depression.

Opioids mainly target the μ-opioid receptor (μOR) in neuronal membranes. Activation of the receptor modulates the behaviour of several membrane ion channels along the nociceptive pathways — those neuronal pathways in the nervous system that respond exclusively to painful or potentially painful stimuli — and in central pain-processing centres. But the side effects associated with these drugs include respiratory depression, constipation and addiction. The development of safe, abuse-free opioid analgesics therefore represents a long-standing scientific challenge[3].

Several strategies have been used to try to develop analgesics that activate the μOR but which are free from adverse side effects. These include the development of partial activators of the μOR (ref. 4); of compounds that activate the μOR but block other opioid-receptor subtypes[5]; and of compounds that are restricted to the peripheral, rather than the central, nervous system[6]. Other approaches have involved molecules that bind to the μOR only in acidic environments[7] (which are often associated with damaged tissue), and compounds that target opioid receptors formed from more than one subtype[8]. A more-recent strategy has been to make allosteric modulators of the μOR — compounds that activate it by binding to a region on the receptor other than its active site[9]. Each of these approaches has enhanced our understanding of the μOR system, but has not led to the identification of a safe analgesic.

In the past few years, it has become increasingly evident that different ligand molecules can activate receptors in different ways[10]. In the case of G-protein-coupled receptors (the family of receptors to which the μOR belongs), some ligands selectively activate a signalling pathway that involves the eponymous G protein, whereas others activate signalling through a protein called β-arrestin-2. This selectivity is called biased agonism.

Morphine has no bias for the G-protein or β-arrestin-2 signalling pathways. In 1999, a study[11] showed that the analgesic effect of morphine is enhanced in mice that lack β-arrestin-2 compared to the effect in wild-type mice, and that several of the drug's side effects were reduced. Other studies have since shown similar effects of morphine in mice in which the expression of β-arrestin-2 has been downregulated or inhibited[12,13]. These findings support the idea[14] that opioid agonists with a strong bias towards G-protein-mediated signalling will retain their analgesic properties, but produce fewer side effects than unbiased opioids. Several laboratories have since identified G-protein-biased μOR agonists[4,15–19], many of which clearly separate analgesia from adverse side effects. One of these compounds, known as TRV130, is currently in phase III clinical trials[16]. But how well the G-protein bias of these compounds correlates with analgesic efficacy and the reduction of unwanted side effects is not clear.

Schmid *et al.* tackle the issue of biased opioid signalling head on, and suggest a way to develop analgesic opioids that do not cause respiratory depression. The authors used a previously unreported, selective μOR agonist as a starting point, and synthesized analogues of the compound that contained modifications in two regions of the molecule. They then assessed the analogues *in vitro* for signalling bias, and *in vivo* for analgesic efficacy and effects on respiratory depression in mice. All of the compounds were found to be μOR agonists and produced μOR-dependent analgesia.

The authors observed a robust relationship between *in vitro* G-protein bias and *in vivo* analgesia and respiratory depression: compounds that exhibited higher G-protein bias were stronger pain relievers and caused less respiratory depression. For example, they observed that fentanyl and one of their new compounds (SR-11501; Fig. 1a) exhibit arrestin bias and have a narrow, threefold therapeutic window — that is, the dose of the compounds at which analgesia occurs is approximately three times lower than the dose at which respiratory depression occurs. By contrast, another compound (SR-17018; Fig. 1b) has a robust G-protein bias and a more than 25-fold therapeutic window.

Schmid and colleagues report that SR-17018 does not activate the β-arrestin-2 signalling pathway *in vitro* even at high concentrations, and does not block arrestin recruitment by classic μOR agonists. These results suggest that SR-17018 stabilizes a conformation of μOR that has no affinity for β-arrestin-2, supporting the idea that the pharmacological effects of SR-17018 are attributable solely to its G-protein bias. This distinguishes SR-17018 from previously reported μOR ligands.

These results are exciting, but it remains to be seen whether G-protein-biased compounds could be developed as non-addictive opioid analgesics. A few promising drug candidates have been identified that might help to answer this question — for example, in mice, the preclinical candidates PZM21 (ref. 17) and mitragynine pseudoindoxyl[19] exhibit some G-protein bias and clearly separate μOR-mediated analgesia from adverse side effects, including respiratory depression, constipation and capacity for abuse. However, a study in rodents indicates that TRV130 retains the potential for abuse[20].

Nevertheless, compounds that have similar properties to those identified by Schmid and colleagues are strong candidates for the development of truly safe analgesics. Rational approaches to the design of such a drug will require the identification of the amino-acid residues in the μOR's binding pocket that are responsible for biased agonism. A long-term goal must therefore be to generate crystals of the receptor in its G-protein-biased and arrestin-biased conformations, so that the structures can be solved using X-ray crystallography. The compounds identified by Schmid *et al.* should also inform our understanding of signalling through G-protein-coupled receptors in general. Given that such receptors are implicated in many diseases, this could pave the way for the development of numerous drugs that have minimal side effects. ■

Susruta Majumdar *is in the Department of Neurology, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA.* **Lakshmi A. Devi** *is in the Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA.*
e-mails: majumdas@mskcc.org; lakshmi.devi@mssm.edu

1. Schmid, C. L. *et al. Cell* **171,** 1165–1175 (2017).
2. Institute of Medicine. *Relieving Pain in America* (Natl Acad. Press, 2011).
3. Pasternak, G. W. & Pan, Y.-X. *Pharmacol. Rev.* **65,** 1257–1317 (2013).
4. Grinnell, S. G. *et al. Synapse* **70,** 395–407 (2016).
5. Ananthan, S. *AAPS J.* **8,** E118–E125 (2006).
6. Eans, S. O. *et al. J. Med. Chem.* **58,** 4905–4917 (2015).
7. Spahn, V. *et al. Science* **355,** 966–969 (2017).
8. Fujita, W., Gomes, I. & Devi, L. A. *Br. J. Pharmacol.* **172,** 375–387 (2015).
9. Burford, N. T. *et al. Proc. Natl Acad. Sci. USA* **110,** 10830–10835 (2013).
10. Rankovic, Z., Brust, T. F. & Bohn, L. M. *Bioorg. Med. Chem. Lett.* **26,** 241–250 (2016).
11. Bohn, L. M. *et al. Science* **286,** 2495–2498 (1999).
12. Bu, H., Liu, X., Tian, X., Yang, H. & Gao, F. *Int. J. Neurosci.* **125,** 56–65 (2015).
13. Li, Y. *et al. Int. J. Mol. Sci.* **10,** 954–963 (2009).
14. Raehal, K. M., Walker, J. K. & Bohn, L. M. *J. Pharmacol. Exp. Ther.* **314,** 1195–1201 (2005).
15. Harding, W. W. *et al. J. Med. Chem.* **48,** 4765–4771 (2005).
16. DeWire, S. M. *et al. J. Pharmacol. Exp. Ther.* **344,** 708–717 (2013).
17. Manglik, A. *et al. Nature* **537,** 185–190 (2016).
18. Kruegel, A. C. *et al. J. Am. Chem. Soc.* **138,** 6754–6764 (2016).
19. Váradi, A. *et al. J. Med. Chem.* **59,** 8381–8397 (2016).
20. Altarifi, A. A. *et al. J. Psychopharmacol.* **31,** 730–739 (2017).

GLOBAL WARMING

Homing in on a key factor of climate change

The sensitivity of Earth's climate to atmospheric carbon dioxide levels is a big unknown in predicting future global warming. A compelling analysis suggests that we can rule out high estimates of this sensitivity. SEE LETTER P.319

PIERS FORSTER

The quantity known as equilibrium climate sensitivity is crucial for understanding Earth's future temperature[1], and ongoing uncertainty about its value makes it harder to adequately prepare for the long-term effects of climate change[2]. This key parameter enumerates the increase in Earth's average surface temperature that would occur if atmospheric carbon dioxide concentrations were doubled and the climate system was given enough time to reach an equilibrium state. More than 150 estimates of equilibrium climate sensitivity (ECS) have been published[3], many of which suggest that worryingly high sensitivities are possible — including one that was published in *Nature* just a few weeks ago[4]. On page 319, Cox *et al.*[5] use an ingenious approach to rule out high estimates. If correct, this would improve the chances of achieving internationally agreed targets for minimizing global warming.

The measurements of many different properties, such as the height of Everest or the speed of light, have often been refined. This has helped to bring certainty to science and thereby driven progress. But ECS has not capitulated to these scientific norms and remains stubbornly uncertain. It has also become a focus for those who doubt the robustness of climate science, who use it to suggest that the field as a whole is intrinsically unreliable. Despite the huge progress in our understanding of climate science over the past 40 years, the Intergovernmental Panel on Climate Change (IPCC) concluded[1] in 2013 that there is a 66% likelihood of ECS being between 1.5 °C and 4.5 °C (Fig. 1). This is little different from the range first postulated[6] by the meteorologist Jule Charney and colleagues in 1979.

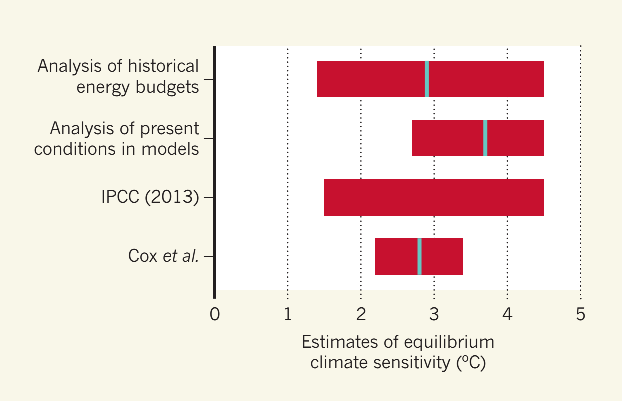


Figure 1 | Estimates of equilibrium climate sensitivity (ECS). ECS quantifies the increase in Earth's average surface temperature that would occur if atmospheric carbon dioxide levels were doubled and the climate system was allowed to reach an equilibrium state. Estimates of ECS vary depending on the evidence used (such as records of Earth's energy budget[9] and analyses[4] of present climate conditions produced by models). The estimate[1] from the Intergovernmental Panel on Climate Change (IPCC) published in 2013 is based on several lines of evidence. Cox *et al.*[5] now report estimates based on an analysis of surface-temperature variation predicted by climate models. Their analysis rules out high estimates of ECS. Bars depict ranges for which there is a 66% likelihood of the value being correct; for the top two bars, these ranges have been inferred from the data in references 4 and 9. Best estimates of ECS for each range, if available, are indicated by a blue line.

Cox and co-workers' estimate is exciting because it develops an underexplored line of evidence: the natural variability of global temperature. The authors also provide the first convincing evidence that we are not living in a world in which ECS is greater than the range of values thought likely by the IPCC. This is important, because estimates of ECS based on the historical temperature record have largely been unable to exclude high values that would invariably result in world-devastating warming of 4 °C or more by 2100.

Past research that seemingly constrained the top end of ECS estimates to lower values often excluded major uncertainties, or worked from a previous estimate of ECS that was skewed towards low values. The published ranges therefore depended on the researchers' assumptions about ECS, rather than the evidence. By contrast, Cox *et al.* started from climate-model values that are at the upper end of the IPCC range, and used evidence to effectively rule out catastrophically high values: they estimate that there is a 66% likelihood of ECS being between 2.2 °C and 3.4 °C, with less than a 1% chance of it being greater than 4.5 °C (Fig. 1).

The idea underpinning this work is so enviably simple that it will make climate scientists ask, "Why didn't I think of that?" The authors examined the variability of surface temperature in terms of its variance and autocorrelation — the 'memory' of a previous year's surface temperature that is retained in measurements taken the following year. They then developed a theory-derived metric of surface-temperature variability and evaluated this metric in historical simulations from 22 computational models of the Earth system, ultimately finding that it is a good predictor of the inherent ECS of each of the models.

Cox *et al.* then used the relationship between the metric and the ECS found in the models as a constraint on ECS in the real world. Their analysis revealed that only climate models that produce relatively small values of ECS match the variability seen in the historical temperature record. It turns out that, in general, climate models have considerable memory in their climate systems, so if one year is abnormally hot, for example, then the next year is likely also to be hot. The historical temperature record, however, does not seem to have as much system memory as most models. This means that some models have both autocorrelations and ECS values that are too high.

These new findings must be interpreted carefully. ECS is arguably the main factor that governs uncertainty in projected temperatures, but is not the only factor. For example, Earth-system feedbacks such as the effects of permafrost melting are expected to increase warming. Climate models often exclude these feedbacks, reducing the projected warming. In models that have an ECS that is too high, such exclusions could potentially compensate for the effects of the inflated ECS value.

It is also crucial to examine other lines of evidence when assessing ECS. The best estimates of ECS that have been made by analysing Earth's energy budget (the balance of the energy received by Earth from the Sun and the energy radiated back to space) are relatively low, at around 2 °C (ref. 7). But recent work[8] is helping us to understand that ECS values inferred from energy-budget changes over the past century are probably low, and shows that a higher value is more applicable when projecting future change. Applying such a correction to the original estimates[9] brings their values very much in line with Cox and co-workers' estimate (Fig. 1).

By contrast, analyses[3] of present climate conditions (particularly cloud properties) produced by models show that the models that best represent today's climate have ECS values greater than 3 °C. Indeed, one of the most recent of these analyses[4] showed that models with an ECS of around 4 °C best captured today's climate across nine emergent constraints (Fig. 1). In my view, Cox and colleagues' estimate and the estimates produced by analysing the historical energy budget carry the most weight, because they are based on simpler physical theories of climate forcing and response, and do not directly require the use of a climate model that correctly represents cloud. To resolve which estimates are most accurate, more research is needed to compare the different lines of evidence and to improve the representation of clouds in models.

> ***"The idea underpinning this work is so enviably simple that it will make climate scientists ask, "Why didn't I think of that?"***

I hope that a much more refined estimate of ECS can be made from the different lines of evidence by the time the next IPCC assessment is published in 2021. If the upper limit of ECS can truly be constrained to a lower value than is currently expected, then the risk of very high surface-temperature changes occurring in the future will decrease. This, in turn, would improve the chances of keeping the temperature increase well below 2 °C above pre-industrial levels, the target of the Paris Agreement under the United Nations Framework Convention on Climate Change. So, rather than be jealous, I should thank Cox and colleagues for helping me to sleep a little easier in my bed at night. ■

Piers Forster *is at the School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK.*
e-mail: p.m.forster@leeds.ac.uk

1. Stocker, T. F. *et al.* (eds) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, 2013).
2. Weitzman, M. L. *Rev. Environ. Econ. Policy* **5,** 275–292 (2011).
3. Knutti, R., Rugenstein, M. A. A. & Hegerl, G. C. *Nature Geosci.* **10,** 727–736 (2017).
4. Brown, P. T. & Caldeira, K. *Nature* **552,** 45–50 (2017).
5. Cox, P. M., Huntingford, C. & Williamson, M. S. *Nature* **553,** 319–322 (2017).
6. Charney, J. G. *et al. Carbon Dioxide and Climate: A Scientific Assessment* (Natl Acad. Sci., 1979).
7. Forster, P. M. *Annu. Rev. Earth Planet. Sci.* **44,** 85–106 (2016).
8. Ceppi, P. & Gregory, J. M. *Proc. Natl Acad. Sci. USA* **114,** 13126–13131 (2017).
9. Armour, K. C. *Nature Clim. Change* **7,** 331–335 (2017).

MOLECULAR BIOLOGY

Limitless translation limits translation

Evidence has now been found that ribosomes — the cell's translational apparatus — can pass beyond the main protein-coding region of messenger RNAs to form 'traffic jams' that inhibit protein expression. SEE LETTER P.356

PETRA VAN DAMME

During the process of translation, molecular machines in the cell called ribosomes use sequences encoded by messenger RNAs as templates for protein synthesis. On page 356, Yordanova *et al.*[1] propose an intriguing mechanism that might limit the number of protein molecules that can be synthesized from a single mRNA. It involves the formation of a queue of ribosomes on the mRNA, downstream of the main protein-coding region.

The conventional view of translation in eukaryotes — organisms such as fungi, plants and animals — is that each mRNA consists of a stretch of nucleotides that contains an open reading frame (ORF), which encodes a single protein containing more than 100 amino-acid residues. But over the past decade, the advent of technologies such as ribosome profiling[2] has revealed that a more-diverse range of ORF sequences can, in fact, be translated. For example, numerous small upstream ORFs (uORFs) have been identified whose translation might regulate expression of the main ORF.

Ribosome profiling has also revealed a wealth of events in which translation is initiated at alternative start codons[3] (triplets of nucleotides other than the triplets at which translation is normally assumed to initiate), and read-through events[4] in which translation continues beyond the stop codon (the nucleotide triplet at the end of the ORF). Not only do these two types of event increase the overall diversity of proteoforms (molecular forms of proteins produced from genes)[5], but they have also emerged as regulatory mechanisms for hundreds of genes in eukaryotic genomes. Other regulatory mechanisms for translation are also known, including ribosome stalling, in which obstacles impede ribosome movement along mRNAs.

Yordanova *et al.* now propose another evolutionarily conserved mechanism for translational control. They suggest that sporadic stop-codon read-throughs can lead to the formation of ribosome queues at downstream stalling sites, such that the queue length is proportional to the number of protein molecules that have been synthesized. The authors define the region between the end of the main ORF and the next in-frame stop codon (that is, the next nucleotide triplet that would be recognized as a stop codon by a

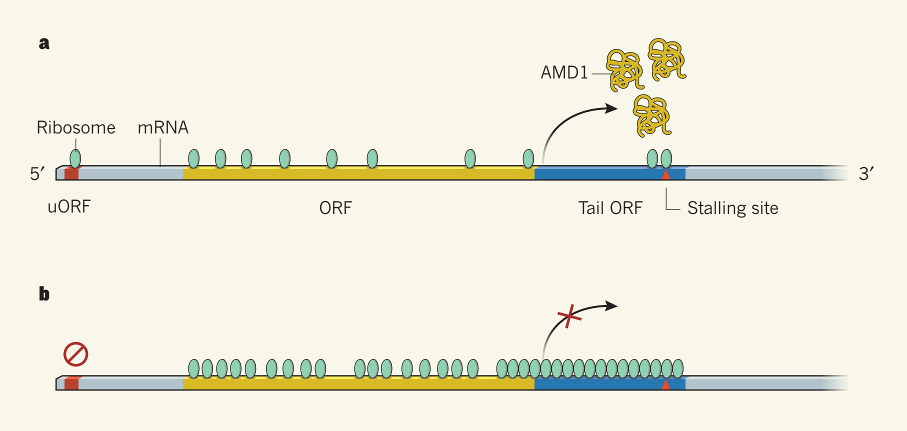


Figure 1 | A proposed regulatory mechanism for translation of the *AMD1* gene. **a**, The messenger RNA for the AMD1 protein contains an open reading frame (ORF) that encodes the protein's amino-acid sequence, and also an upstream open reading frame (uORF), the translation of which regulates translation of the ORF (ref. 6). Grey regions of mRNA are not translated. Yordanova *et al.*[1] propose that ribosomes (the cellular machinery responsible for translation) can sporadically enter and translate a region called the tail ORF, rather than stopping at the end of the main ORF. The ribosomes eventually halt at a stalling site (a nucleotide sequence that halts translation) in the tail ORF. **b**, When uORF-mediated regulation is blocked, translation initiation at the main ORF increases, so that ribosomes accumulate more quickly in the tail ORF, forming a queue. When the queue extends beyond the tail ORF, translation of the main ORF is impaired, limiting the number of protein molecules that can be synthesized from a single *AMD1* mRNA template.

ribosome translating beyond the main ORF's stop codon) as the tail ORF. They suggest that translation is halted when queuing ribosomes in the tail ORF extend into the main ORF (Fig. 1).

The authors were inspired to propose this mechanism after inspecting publicly available ribosome-translation profiles for a protein called adenosylmethionine decarboxylase 1 (encoded by the *AMD1* gene), the translation of which is tightly controlled. The profiles revealed translation of a uORF in the *AMD1* mRNA, as previously reported[6], but also a high density of ribosomes in a region known as the 3′ trailer of the mRNA, downstream of the *AMD1* stop codon. This suggested that a stop-codon read-through had occurred, allowing ribosomes to accumulate in the tail ORF of *AMD1*.

Yordanova and colleagues performed experiments showing that stable peptidyl–transfer RNA complexes (which form between tRNA and the nascent protein chain during translation) are generated when tail-ORF sequences are translated, and that complex formation occurs before translation reaches the stop codon at the end of the tail ORF. This confirmed that the proposed read-through could occur, and that translation could stall in the tail ORF. The authors also constructed a mutant mRNA in which the *AMD1* stop codon was replaced by a sense codon (a nucleotide triplet that encodes an amino acid), in the expectation that translation would occur uninterrupted through the mutated sequence. However, almost no *AMD1* translation occurred with this mutant — the expected extended proteoform was produced in nearly undetectable levels.

To explore the mechanisms that affect the levels of expressed protein, Yordanova *et al.* used a strategy[7] known as StopGo, which allows the cleavage and release of nascent protein chains at chosen positions during translation, but then allows ribosomes to resume translation of the downstream sequence. The authors used StopGo to cleave nascent proteins before translation of the *AMD1* stop codon in the wild-type mRNA, and before translation of the sense codon in the mutant mRNA. They observed that the amount of AMD1 protein subsequently produced from the mutant mRNA was lower than the amount produced from the wild-type mRNA — even though the amino-acid sequences of the proteins were identical.

This result suggests that the tail ORF must lower protein expression by influencing translation, rather than by reducing the stability of the produced protein. The finding is at odds with previous work[8] showing that proteins are generally destabilized when their sequences are extended by a stop-codon read-through. Yordanova and colleagues' experimental data collectively show that the effects of translation of the *AMD1* tail ORF are independent of the main coding sequence, mRNA stability, common protein-degradation and cleavage pathways, or whether the expressed protein is secreted by cells.

The findings are therefore consistent with the idea that translation is halted when a ribosome queue in the tail ORF extends into the main *AMD1* coding region. Note that such regulation could work only if *AMD1* translation is dysregulated and high, and would not apply under standard conditions. The authors also observed that translational output was more strongly reduced in experiments that increased the efficiency of read-throughs, thereby accelerating queue formation — in agreement with the model.

A note of caution is warranted, because Yordanova and co-workers have not directly observed long ribosome queues. The proposed ribosome stalling might also occur only transiently, thereby increasing the time required to attain full ribosome coverage of the tail ORF and so decreasing the overall impact of ribosome-queue formation on translation. Furthermore, besides the experimentally validated traffic-jam model[9] (in which ribosomes collide and form queues, blocking translation initiation), other models for how ribosome stalling interferes with translation have been proposed. For example, it has been suggested that stalled ribosomes fall off mRNA following collision with a trailing ribosome; this model conflicts with the idea that long ribosome queues could form[10]. Peculiarities of the expression systems used by Yordanova *et al.* might also underlie some of the authors' observations. Finally, their data do not rule out the possible involvement of other factors that could cause the downregulation of protein expression, such as protein-degradation pathways that occur at the same time as translation.

Nevertheless, the authors' findings will surely inspire future endeavours to obtain concrete proof of the proposed mechanism, and to assess how widely it is used to limit protein synthesis. Single-molecule imaging of translation on individual mRNA molecules, in real time and in live cells, might eventually allow simultaneous observation of mRNAs and their protein products[11]. ■

Petra Van Damme *is at the VIB-UGent Center for Medical Biotechnology, and in the Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium.*
e-mail: petra.vandamme@vib-ugent.be

1. Yordanova, M. M. *et al. Nature* **553,** 356–360 (2018).
2. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. *Science* **324,** 218–223 (2009).
3. Van Damme, P., Gawron, D., Van Criekinge, W. & Menschaert, G. *Mol. Cell. Proteomics* **13,** 1245–1261 (2014).
4. Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R. & Weissman, J. S. *eLife* **2,** e01179 (2013).
5. Smith, L. M., Kelleher, N. L. & The Consortium for Top Down Proteomics *Nature Methods* **10,** 186–187 (2013).
6. Law, G. L., Raney, A., Heusner, C. & Morris, D. R. *J. Biol. Chem.* **276,** 38036–38043 (2001).
7. Doronina, V. A. *et al. Mol. Cell. Biol.* **28,** 4227–4239 (2008).
8. Arribere, J. A. *et al. Nature* **534,** 719–723 (2016).
9. MacDonald, C. T., Gibbs, J. H. & Pipkin, A. C. *Biopolymers* **6,** 1–25 (1968).
10. Ferrin, M. A. & Subramaniam, A. R. *eLife* **6,** e23629 (2017).
11. Wang, C., Han, B., Zhou, R. & Zhuang, X. *Cell* **165,** 990–1001 (2016).

This article was published online on 3 January 2018.

ARTICLE

doi:10.1038/nature25178

Dietary trehalose enhances virulence of epidemic *Clostridium difficile*

J. Collins[1], C. Robinson[2], H. Danhof[1], C. W. Knetsch[3], H. C. van Leeuwen[3], T. D. Lawley[4], J. M. Auchtung[1] & R. A. Britton[1]

Clostridium difficile disease has recently increased to become a dominant nosocomial pathogen in North America and Europe, although little is known about what has driven this emergence. Here we show that two epidemic ribotypes (RT027 and RT078) have acquired unique mechanisms to metabolize low concentrations of the disaccharide trehalose. RT027 strains contain a single point mutation in the trehalose repressor that increases the sensitivity of this ribotype to trehalose by more than 500-fold. Furthermore, dietary trehalose increases the virulence of a RT027 strain in a mouse model of infection. RT078 strains acquired a cluster of four genes involved in trehalose metabolism, including a PTS permease that is both necessary and sufficient for growth on low concentrations of trehalose. We propose that the implementation of trehalose as a food additive into the human diet, shortly before the emergence of these two epidemic lineages, helped select for their emergence and contributed to hypervirulence.

Whole-genome sequencing analysis of *C. difficile* ribotype 027 (RT027) strains demonstrated that two independent lineages emerged in North America from 2000 to 2003 (ref. 1). Comparison with historic, pre-epidemic, RT027 strains showed that both epidemic lineages acquired a mutation in the *gyrA* gene, leading to increased resistance to fluoroquinolone antibiotics. While the development of fluoroquinolone resistance has almost certainly played a role in the spread of RT027 strains, fluoroquinolone resistance has also been observed in non-epidemic *C. difficile* ribotypes and identified in strains dating back to the mid-1980s[2,3]. Thus, other factors probably contributed to the emergence of epidemic RT027 strains.

The prevalence of a second *C. difficile* ribotype, RT078, increased tenfold in hospitals and clinics from 1995 to 2007 and was associated with increased disease severity[4]. However, the mechanisms responsible for increased virulence remain unknown[5–8]. It is noteworthy that RT027 and RT078 lineages are phylogenetically distant from one another (Extended Data Fig. 1), indicating that the evolutionary changes leading to concurrent increases in epidemics and disease severity might have emerged by independent mechanisms[9].

RT027 and RT078 strains grow on low trehalose

Ribotype 027 strains exhibit a competitive advantage over non-RT027 strains *in vitro* and in mouse models of *C. difficile* infection[10]. To investigate potential mechanisms for increased fitness, we examined carbon source utilization in an epidemic RT027 isolate (CD2015) using Biolog 96-well Phenotype MicroArray carbon source plates (see Methods and Extended Data Table 1). Out of several carbon sources identified that supported CD2015 growth, we found the disaccharide trehalose increased the growth yield of CD2015 by approximately fivefold compared with a non-RT027 strain. To examine the specificity of enhanced growth on trehalose across *C. difficile* lineages, 21 strains encompassing 9 ribotypes were grown on a defined minimal medium (DMM) supplemented with glucose or trehalose as the sole carbon source. All *C. difficile* strains grew robustly with 20 mM glucose; however, only epidemic RT027 ($n=8$) and RT078 ($n=3$) strains exhibited enhanced growth on an equivalent trehalose concentration (10 mM; Fig. 1). Increasing the trehalose concentration to 50 mM enabled growth in most ribotypes (Extended Data Fig. 2a).

Molecular basis for RT027 growth on low trehalose

To identify the genetic basis for enhanced trehalose metabolism, we compared multiple *C. difficile* genomes. All *C. difficile* genomes encode a putative phosphotrehalase enzyme (TreA) preceded by a transcriptional repressor (TreR) (Fig. 2a). Phosphotrehalase enzymes metabolize trehalose-6-phosphate into glucose and glucose-6-phosphate. To test whether *treA* was essential for trehalose metabolism, we generated *treA* deletion mutants in the RT027 strain R20291 (R20291Δ*treA*) and the RT012 strain CD630 (CD630Δ*treA*) and grew them in DMM supplemented with 50 mM trehalose. The lack of *treA* prevented growth in both knockout strains, which could be complemented by plasmid expression of *treA* (Extended Data Fig. 2b). Thus, *treA* is required to metabolize trehalose.

We next asked whether RT027 strains have altered regulation of the *treA* gene compared with other ribotypes. To test this hypothesis and determine the minimum level of trehalose required to activate *treA* expression, we grew CD2015 (RT027) and CD2048 (RT053) and exposed them to increasing amounts of trehalose. We found that the

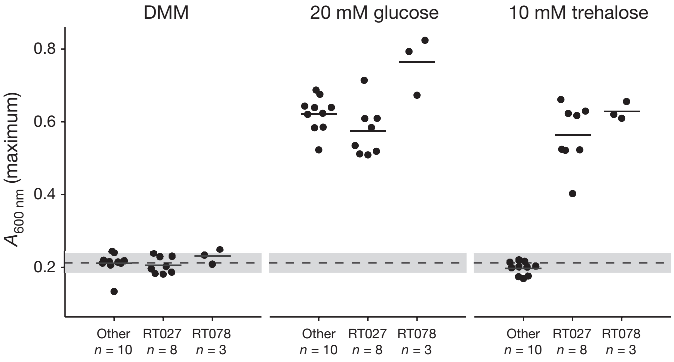


Figure 1 | Only RT027 and RT078 strains show enhanced growth on 10 mM trehalose. Dashed grey line and band indicate mean growth and s.d. in DMM without a carbon source for all samples ($n=21$). Solid lines are mean growth yield (absorbance at 600 nm, $A_{600\ \mathrm{nm}}$) for groups: non-RT027/078 ($n=10$), RT027 ($n=8$), and RT078 ($n=3$). All points represent biologically independent samples.

[1]Baylor College of Medicine, Department of Molecular Virology and Microbiology, One Baylor Plaza, Houston, Texas 77030, USA. [2]University of Oregon, Institute for Molecular Biology, 1318 Franklin Boulevard, Eugene, Oregon 97403, USA. [3]Leiden University Medical Centre, Department of Medical Microbiology, Albinusdreef 2, 2333 ZA Leiden, The Netherlands. [4]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK.

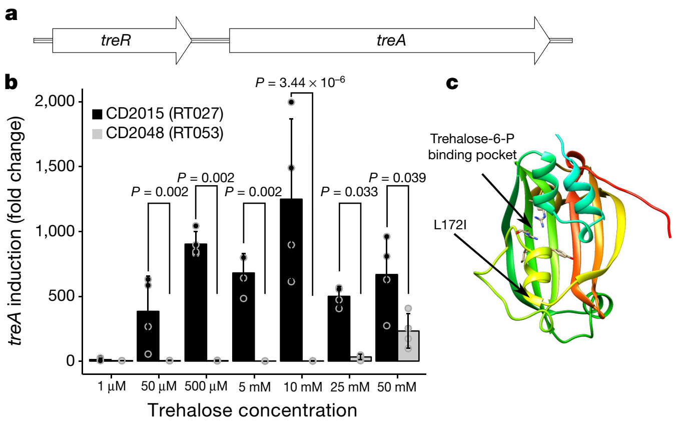


Figure 2 | The *treA* gene is responsible for trehalose metabolism. **a**, Trehalose metabolism operon found in all *C. difficile* strains, consisting of a phosphotrehalase (*treA*) and its transcriptional regulator (*treR*). **b**, RT027 strains strongly induce *treA* at 50 µM trehalose and at a significantly higher level than non-RT027 strains ($n = 4$ biologically independent samples per trehalose concentration/strain). Bars are average fold increase, error bars are s.d.; P values derived from t-test (two-tailed) and Holm corrected for multiple comparisons. **c**, Structure of TreR monomer highlighting proximity of L172I mutation to trehalose-6-P binding pocket.

RT027 strain turned on *treA* expression at 50 µM trehalose, a concentration 500-fold lower than that required to turn on *treA* in RT053 (Fig. 2b). To confirm this phenotype, we took four RT027 strains and four non-RT027 strains and measured expression of *treA* in a single trehalose concentration. Again, RT027 strains exhibited significantly higher *treA* expression than all other ribotypes ($P = 0.029$; Extended Data Fig. 3). These results support the idea that RT027 strains are exquisitely sensitive to low concentrations of trehalose.

Sequence alignment of the trehalose operon across 1,010 sequenced *C. difficile* strains revealed a conserved single nucleotide polymorphism (SNP) within the *treR* gene of all RT027 strains ($\text{TreR}_{\text{RT027}}$) (Extended Data Fig. 4a). The SNP encodes an L172I amino acid substitution near the predicted effector (trehalose-6-phosphate) binding pocket of TreR (Fig. 2c), a site that is highly conserved across multiple species (93.9% conservation; Extended Data Fig. 4b). This SNP is found not only in every RT027 strain sequenced so far, but also in a newly isolated fluoroquinolone sensitive ribotype (RT244) that has caused community-acquired epidemic outbreaks in Australia[11,12] and other ribotypes very closely related to RT027, such as RT176 which has caused epidemic outbreaks in the Czech Republic and Poland[13,14]. Like RT027 strains, the RT244 strains DL3110 and DL3111 can grow on 10 mM trehalose (Extended Data Fig. 2c).

To determine the types of spontaneous mutation that lead to enhanced trehalose utilization, we cultivated several non-RT027/RT078 strains under low trehalose concentrations in minibioreactors[10]. After 3 days of continuous cultivation, 13 independent spontaneous mutants capable of growing on low concentrations (<10 mM) of trehalose were isolated. All 13 mutants contained either nonsense or missense mutations in the *treR* gene (Extended Data Table 2).

Effect of trehalose metabolism on disease severity

To test whether the ability of *C. difficile* RT027 strains to metabolize trehalose impacts disease severity, we performed two different experiments. In the first, humanized microbiota mice were challenged with 10^4 spores of either R20291 (RT027, $n = 27$) or R20291Δ*treA* ($n = 28$). After infection, trehalose (5 mM) was provided *ad libitum* in the drinking water and disease progression monitored. The R20291Δ*treA* mutant demonstrated a marked decrease in mortality (33.3% versus 78.6%) compared with R20291 (78% lower risk with R20291Δ*treA*; hazard ratio 0.22; 95% confidence interval 0.09–0.59; $P = 0.003$, likelihood ratio test $P = 0.002$; Fig. 3a). In the second experiment, we infected two groups of humanized microbiota mice with RT027 strain R20291.

One group received 5 mM trehalose in water as well as a daily gavage of 300 mM trehalose ($n = 28$) to mimic a dose expected in a meal for humans, whereas the control group ($n = 27$) received a water control. Trehalose addition was found to cause increased mortality compared with the RT027-infected mice without dietary trehalose (threefold increased risk with trehalose; hazard ratio, 3.20; 95% confidence interval 1.09–9.42; $P = 0.035$, likelihood ratio test $P = 0.026$; Fig. 3b). Combined, these results show that metabolism of dietary trehalose can contribute to disease severity of RT027 *C. difficile* strains.

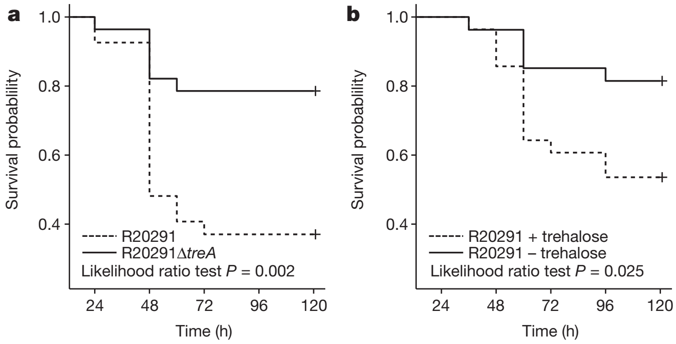


Figure 3 | Trehalose metabolism increases virulence. **a**, Mice infected with R20291Δ*treA* ($n = 27$ animals) have significantly attenuated risk of mortality compared with mice infected with R20291 ($n = 28$ animals) (78% lower risk with Δ*treA* mutant; hazard ratio 0.22; 95% confidence interval 0.09–0.59; $P = 0.003$). **b**, Mice infected with R20291 (RT027) have a significantly higher risk of mortality when trehalose is supplemented in the diet ($n = 28$ animals) than those with no trehalose supplementation ($n = 27$ animals) (threefold increased risk with trehalose; hazard ratio 3.20; 95% confidence interval 1.09–9.42; $P = 0.035$). Experiments were repeated twice. All statistical tests were two-sided.

To identify the cause of increased disease severity when trehalose is present, we challenged mice with either R20291 or R20291Δ*treA* and provided 5 mM trehalose *ad libitum* in the drinking water. Forty-eight hours after challenge, *C. difficile* load and toxin levels were measured. Over two independent experiments, no significant difference in *C. difficile* numbers was observed; however, a significant increase in the relative levels of toxin B was detected (median 9.2×10^4, interquartile range (IQR) 5.1×10^4 to 1.0×10^5 versus median 4.1×10^4, IQR 2.3×10^4 to 4.6×10^4, $P = 0.0268$; Extended Data Fig. 5). This increased toxin production could contribute to increased disease severity.

Molecular basis for RT078 growth on low trehalose

Unlike RT027, RT078 strains do not possess the TreR L172I substitution or other conserved SNPs in the *treRA* operon. To identify sequences of potential relevance to trehalose metabolism, we performed whole-genome comparisons. A four-gene insertion was found in all RT078 strains sequenced so far, annotated to encode a second copy of a phosphotrehalase (TreA2, sharing 55% amino acid identity with TreA), a potential trehalose specific PTS system IIBC component transporter (PtsT), a trehalase family protein that is a putative glycan debranching enzyme (TreX), and a second copy of a TreR repressor protein (TreR2, sharing 44% amino acid identity with TreR) (see Fig. 4a). Genomic comparison of publicly available *C. difficile* genomes revealed the four-gene insertion was present in RT078 and the closely related RT033, RT045, RT066 and RT126 ribotypes and absent from reference genomes of any other *C. difficile* lineage (Extended Data Fig. 6).

To test whether the newly acquired transporter (*ptsT*) was responsible for enhanced trehalose metabolism, a *ptsT* deletion mutant was constructed in a RT078 (CD1015) strain. This strain was unable to grow on DMM supplemented with 10 mM trehalose (Fig. 4b), but retained the ability to grow in medium supplemented with 50 mM trehalose (Extended Data Fig. 2d). The growth defect in this deletion mutant (CD1015Δ*ptsT*) was directly due to the lack of *ptsT* since expression of *ptsT* from an inducible promoter could complement growth on 10 mM trehalose (Fig. 4b).

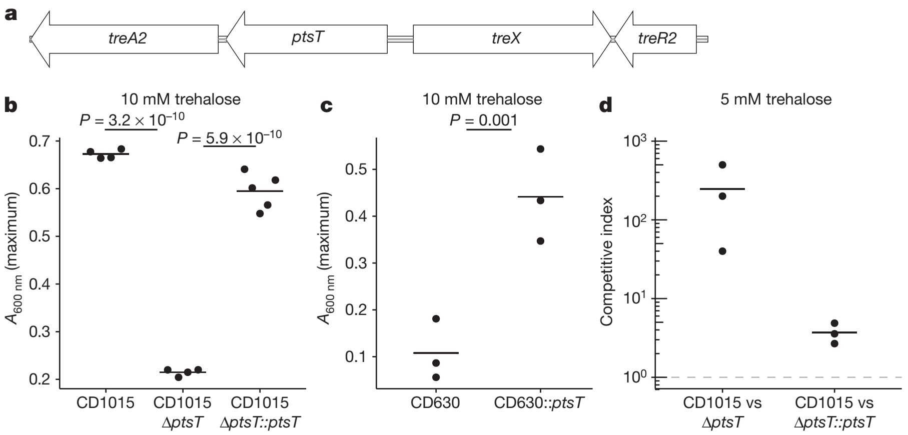


Figure 4 | The *ptsT* gene enables enhanced trehalose metabolism. **a**, Structure of horizontally acquired trehalose metabolism module found in RT078 and closely related strains. **b**, Deletion of the trehalose transporter from a clinical RT078 strain (CD1015) ablates its ability to grow on 10 mM trehalose. Expression of *ptsT* from an inducible plasmid restores growth of CD1015Δ*ptsT* on 10 mM trehalose (CD1015, *n* = 4; CD1015Δ*ptsT*, *n* = 4; CD1015Δ*ptsT*::*ptsT*, *n* = 5). **c**, Expressing *ptsT* from an inducible plasmid enables enhanced growth of CD630 (RT012) on 10 mM trehalose (CD630 *n* = 3; CD630::*ptsT* *n* = 3). **d**, The *ptsT* provides a competitive advantage in complex microbial communities. Dashed grey line (competitive index = 1) indicates equal fitness of the competing strains, points above this line represent out-competition by CD1015. All points (Fig. 4b–d) represent biologically independent samples, bars are mean, *P* values derived from *t*-test (two-tailed) and Holm corrected for multiple comparisons where appropriate.

We next tested whether *ptsT* was sufficient to confer enhanced trehalose utilization in a non-ribotype 078 strain, which fails to grow under low trehalose concentrations. To do this, *ptsT* was expressed from an inducible promoter in strain CD630 (RT012). Expression of *ptsT* was sufficient to allow growth of CD630 in DMM supplemented with 10 mM trehalose (Fig. 4c). Taken together, we conclude that *ptsT* is both necessary and sufficient to support growth on low concentrations of trehalose.

To test whether the expression of *ptsT* could confer a fitness advantage, CD1015 (RT078) was competed against its isogenic CD1015Δ*ptsT* mutant in a human faecal minibioreactor model of *C. difficile* infection[10]. After clindamycin treatment of minibioreactor communities to enable infection, CD1015 and CD1015Δ*ptsT* strains were added together to each reactor and levels monitored over time. Remarkably, the CD1015 strain was found to be significantly more efficient at competing *in vivo* in the presence of a complex microbiota than the CD1015Δ*ptsT* mutant (mean competitive index of 246 on day 7). To ensure the CD1015Δ*ptsT* loss was due to the absence of *ptsT*, CD1015 was competed against the CD1015Δ*ptsT* mutant complemented with *ptsT* from an inducible vector. After 5 days of continuous competition, the wild-type RT078 had a mean competitive index of just 3.7 (Fig. 4d). Hence, *ptsT* provides a competitive fitness advantage to RT078 strains.

Trehalose is observed in the distal gut

Despite the presence of a localized brush border trehalase enzyme in the small intestine, human studies suggest that high levels of trehalose consumption can result in significant amounts reaching the distal ileum and colon[15–17]. To demonstrate that a significant amount of dietary trehalose can survive transit through the small intestine, we gavaged mice with 100 μl (300 mM) trehalose (equivalent to the suggested concentration in ice cream) and measured trehalose levels in the caecum over time. Using clinical *C. difficile* strains as biosensors, we found the level of trehalose to be sufficient to activate *treA* gene expression in the RT027 strain CD2015 but not in RT053 strain CD2048 (Fig. 5a). To test whether we could detect a low dietary amount of trehalose, we gavaged antibiotic-treated mice with 100 μl (5 mM) trehalose and measured *treA* activation in these same strains. Again, the RT027 strain showed significant *treA* activation (Fig. 5b). Finally, to determine whether trehalose is bioavailable in humans at sufficient levels to be used by epidemic *C. difficile* isolates, we tested ileostomy effluent from three anonymous donors consuming their normal diets. In two of three samples, *treA* expression was strongly induced in the RT027 strain CD2015 but not in the RT053 strain CD2048 (Fig. 5c), supporting the notion that levels of trehalose found in food are sufficient to be used by epidemic *C. difficile* strains.

Discussion

Containing an α,α-1,1-glucoside bond between two α-glucose units, trehalose is a non-reducing and extremely stable sugar, resistant both to high temperatures and to acid hydrolysis. Although considered an ideal sugar for use in the food industry, the use of trehalose in the United States and Europe was limited before 2000 owing to the high cost of production (approximately US$700 per kilogram). The innovation of a novel enzymatic method for low-cost production from starch made it commercially viable as a food supplement (approximately US$3 per kilogram)[18]. Granted 'generally recognized as safe' status by the US Food and Drug Administration in 2000 and approved for use in food

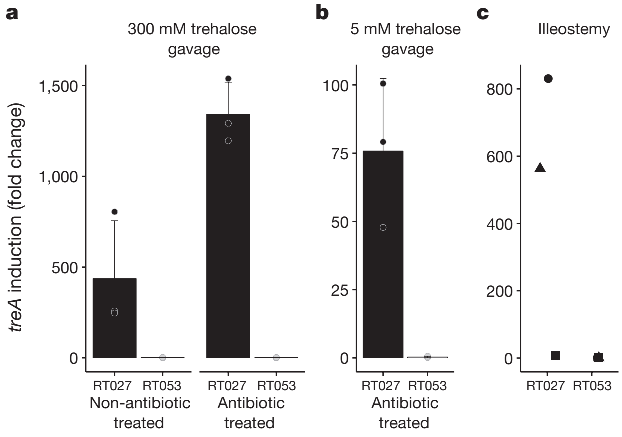


Figure 5 | Trehalose can be detected in mouse caecum and human ileostomy fluid. **a**, Twenty minutes after gavage, trehalose reaches high enough levels in the mouse caecum to turn on expression of *treA* in RT027 but not non-RT027 in both non-antibiotic- and antibiotic-treated mice (*n* = 3 animals per trehalose concentration/strain). **b**, Trehalose can be detected by RT027 but not non-RT027 in the caecum of antibiotic-treated mice gavaged with just 100 μl of 5 mM trehalose (*n* = 3 animals per group). **c**, RT027 strains can detect trehalose in two out of three human ileostomy fluid samples tested from patients eating a normal (no deliberate trehalose addition) diet. Points represent biologically independent replicates, bars are average fold increase, error bars are s.d.

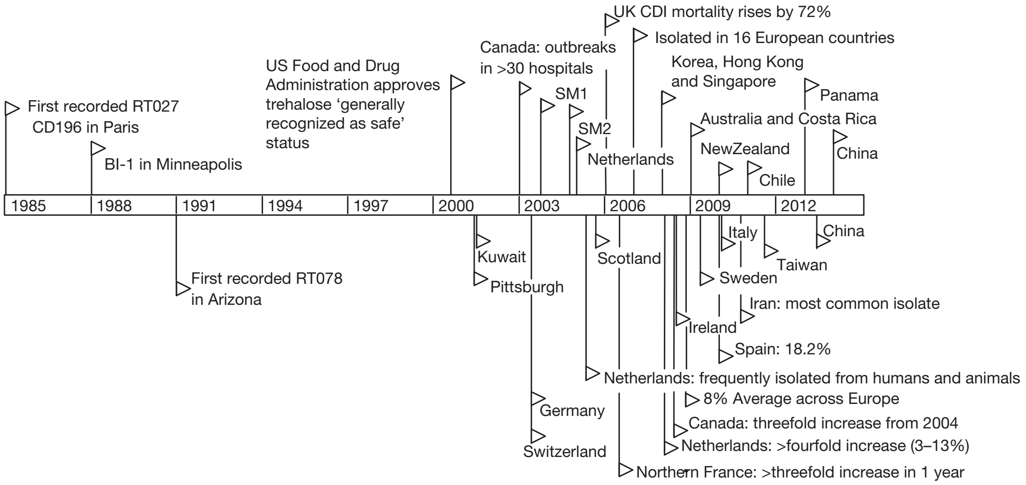

UK CDI mortality rises by 72%
Isolated in 16 European countries
Canada: outbreaks in >30 hospitals
Korea, Hong Kong and Singapore
US Food and Drug Administration approves trehalose 'generally recognized as safe' status
Panama
SM1
First recorded RT027 CD196 in Paris
Australia and Costa Rica
SM2
NewZealand
China
BI-1 in Minneapolis
Netherlands
Chile
1985
1988
1991
1994
1997
2000
2003
2006
2009
2012
Italy
China
Kuwait
Scotland
Taiwan
First recorded RT078 in Arizona
Sweden
Pittsburgh
Iran: most common isolate
Ireland
Spain: 18.2%
Netherlands: frequently isolated from humans and animals
8% Average across Europe
Germany
Canada: threefold increase from 2004
Netherlands: >fourfold increase (3–13%)
Switzerland
Northern France: >threefold increase in 1 year

Figure 6 | Timeline of trehalose adoption and spread of RT027 and RT078 lineages. Flags indicate reported outbreaks or first reports of RT027 (top) or RT078 (bottom) in PubMed. SM1 and SM2, outbreaks at Stoke Mandeville Hospital, Buckinghamshire, UK. CDI, *C. difficile* infection.

in Europe in 2001, reported expected usage ranges from concentrations of 2% to 11.25% for foods including pasta, ground beef, and ice cream. The widespread adoption and use of trehalose in the diet coincides with the emergence of both RT027 and RT078 outbreaks (Fig. 6).

Several lines of evidence support the idea that dietary trehalose has participated in the spread of epidemic *C. difficile* ribotypes. First, the ability of RT027 and RT078 strains to metabolize trehalose was present before epidemic outbreaks. The earliest retrospectively recorded RT027 isolate was the non-epidemic strain CD196, isolated in 1985 in a Paris hospital[19]. Three years later in 1988, another non-epidemic strain RT027 (BI1) was isolated in Minneapolis, Minnesota. Both isolates, in addition to every RT027 strain sequenced so far, contain the L172I substitution in TreR. RT078 strains were also present in humans before 2001, but epidemic outbreaks were not reported until 2003 (ref. 4). Second, RT027 and RT078 lineages are phylogenetically distant clades of *C. difficile*, yet have convergently evolved distinct mechanisms to metabolize low levels of trehalose. Third, increased disease severity of a RT027 strain that can metabolize trehalose in our mouse model of *C. difficile* infection is consistent with increased virulence of RT027 and RT078 ribotypes observed in patients. Fourth, the ability to metabolize trehalose at lower concentrations confers a competitive growth advantage in the presence of a complex intestinal community. Finally, levels of trehalose in ileostomy fluid from patients eating a normal diet are sufficiently high to be detected by RT027 strains. On the basis of these observations, we propose that the widespread adoption and use of the disaccharide trehalose in the human diet has played a significant role in the emergence of these epidemic and hypervirulent strains[20].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 May; accepted 28 November 2017.
Published online 3 January 2018.

1. He, M. *et al.* Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat. Genet.* **45,** 109–113 (2013).
2. Spigaglia, P. *et al.* Fluoroquinolone resistance in *Clostridium difficile* isolates from a prospective study of *C. difficile* infections in Europe. *J. Med. Microbiol.* **57,** 784–789 (2008).
3. Spigaglia, P., Barbanti, F., Dionisi, A. M. & Mastrantonio, P. *Clostridium difficile* isolates resistant to fluoroquinolones in Italy: emergence of PCR ribotype 018. *J. Clin. Microbiol.* **48,** 2892–2896 (2010).
4. Jhung, M. A. *et al.* Toxinotype V *Clostridium difficile* in humans and food animals. *Emerg. Infect. Dis.* **14,** 1039–1045 (2008).
5. Goorhuis, A. *et al.* Emergence of *Clostridium difficile* infection due to a new hypervirulent strain, polymerase chain reaction ribotype 078. *Clin. Infect. Dis.* **47,** 1162–1170 (2008).
6. Gupta, A. & Khanna, S. Community-acquired *Clostridium difficile* infection: an increasing public health threat. *Infect. Drug Resist.* **7,** 63–72 (2014).
7. Limbago, B. M. *et al. Clostridium difficile* strains from community-associated infections. *J. Clin. Microbiol.* **47,** 3004–3007 (2009).
8. Walker, A. S. *et al.* Relationship between bacterial strain type, host biomarkers, and mortality in *Clostridium difficile* infection. *Clin. Infect. Dis.* **56,** 1589–1600 (2013).
9. He, M. *et al.* Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc. Natl Acad. Sci. USA* **107,** 7527–7532 (2010).
10. Robinson, C. D., Auchtung, J. M., Collins, J. & Britton, R. A. Epidemic *Clostridium difficile* strains demonstrate increased competitive fitness compared to nonepidemic isolates. *Infect. Immun.* **82,** 2815–2825 (2014).
11. Lim, S. K. *et al.* Emergence of a ribotype 244 strain of *Clostridium difficile* associated with severe disease and related to the epidemic ribotype 027 strain. *Clin. Infect. Dis.* **58,** 1723–1730 (2014).
12. Eyre, D. W. *et al.* Emergence and spread of predominantly community- onset *Clostridium difficile* PCR ribotype 244 infection in Australia, 2010 to 2012. *Euro Surveill.* **20,** 21059 (2015).
13. Polivkova, S., Krutova, M., Petrlova, K., Benes, J. & Nyc, O. *Clostridium difficile* ribotype 176 – a predictor for high mortality and risk of nosocomial spread? *Anaerobe* **40,** 35–40 (2016).
14. Rupnik, M. *et al.* Distribution of *Clostridium difficile* PCR ribotypes and high proportion of 027 and 176 in some hospitals in four South Eastern European countries. *Anaerobe* **42,** 142–144 (2016).
15. Bergoz, R. Trehalose malabsorption causing intolerance to mushrooms. Report of a probable case. *Gastroenterology* **60,** 909–912 (1971).
16. Bergoz, R., Bolte, J. P. & Meyer zum Bueschenfelde, K.-H. Trehalose tolerance test. Its value as a test for malabsorption. *Scand. J. Gastroenterol.* **8,** 657–663 (1973).
17. Oku, T. & Nakamura, S. Estimation of intestinal trehalase activity from a laxative threshold of trehalose and lactulose on healthy female subjects. *Eur. J. Clin. Nutr.* **54,** 783–788 (2000).
18. Higashiyama, T. Novel functions and applications of trehalose. *Pure Appl. Chem.* **74,** 1263–1269 (2002).
19. Stabler, R. A. *et al.* Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. *Genome Biol.* **10,** R102 (2009).
20. Leffler, D. A. & Lamont, J. T. *Clostridium difficile* infection. *N. Engl. J. Med.* **372,** 1539–1548 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the National Institutes of Health (U01AI124290-01 and 5U19AI09087202). We thank the members of the Ostomy Association of Greater Lansing for anonymously donating samples, and D. Lyras for the RT244 strains. We thank V. Young, D. Mills, J. Walter and M. Costa-Mattioli for comments on the manuscript.

Author Contributions Concept and design of study: R.A.B., J.M.A., J.C. and C.R. Experiments: *C. difficile* growth, J.C. and C.R.; identification of L172I SNP and comparative analysis, C.R. and J.C.; *treA* RT–qPCR, H.D.; mouse infection model, J.C.; genetic manipulation of *C. difficile* strains, J.C. and C.R.; identification of RT078 trehalose insertion, C.W.K., H.C.L. and T.D.L.; faecal minibioreactor competitions, J.M.A.; spontaneous *C. difficile* mutant identification, H.D.; analysis, J.C., C.R., H.D., J.M.A. and R.B. The manuscript was drafted by J.C., J.M.A., and R.A.B., and revised by all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to R.A.B. (robert.britton@bcm.edu).

Reviewer Information *Nature* thanks J. Ballard, E. Pamer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Bacterial strains and growth. A full list of strains can be found in Extended Data Table 3. Carbon source utilization of CD630 (RT012) and CD2015 (clinical RT027) was performed using Biolog Phenotypic Microarray plates. Growth studies were performed under anaerobic conditions (5% hydrogen, 90% nitrogen, 5% carbon dioxide). Strains were cultured overnight in BHI media (Difco) supplemented with 0.5% (w/v) yeast extract. Growth assays used a DMM as described previously[21] supplemented with either trehalose or glucose as indicated. Anhydrous tetracycline was used at 500 ng ml^{-1} to induce expression of *ptsT* or *treA* from ectopic expression vectors.

Comparative genomics. To identify unique functional features in RT078 strains, we reviewed publicly available *C. difficile* genomes covering all phylogenetic lineages[9] using a tool based on BLASTX comparisons of protein annotations[22]. The genomes included in the analysis were PCR RT012 (strain 630, lineage I), RT027 (R20291, lineage II), PCR RT017 (CF5, M68 lineage IV), and RT078 (QCD-23m63, CDM120 lineage V).

Genetic manipulation of *C. difficile*. Inactivation of *treA* in CD630 was accomplished by group-II intron-directed insertion as previously described[23]. Primers were designed to target intron to insert at base pair 177 of *treA* of CD630 (IBS1.2, EBS1, and EBS2; all primers are described in Extended Data Table 3). The resulting *treA* insertion–deletion mutant was verified by PCR using a primer pair (CR064–CR065) designed to flank the *treA* insertion site, resulting in a 350 bp product for the wild-type gene and a 2.4 kbp product for the gene knockout.

Clean deletions in R20291 and CD1015 were performed using a *pyrE* allelic exchange system as described previously[24]. This is the first case of the *pyrE* allelic exchange system being used in the RT078 lineage, which required generation of CD1015Δ*pyrE* before further deletions. Complementation of *treA* and *ptsT* was performed using an anhydrous tetracycline-inducible system as described previously[25]. All plasmid conjugations into *C. difficile* strains were performed with *Escherichia coli* SD46. Cloning was accomplished with a combination of restriction digest and ligase cycling reactions as described previously[26]. Primers and detailed plasmid maps for construction of knockout strains are available at the links provided in Extended Data Table 3.

Quantitative PCR with reverse transcription. Strains were grown overnight and subcultured 1:50 into DMM supplemented with 20 mM succinate. Upon reaching an $A_{600\,\mathrm{nm}}$ of 0.2–0.3, indicated concentrations of trehalose were added to the culture. After incubation for 30 min, *C. difficile* cells were collected by centrifugation, resuspended in RNALater solution (Invitrogen), and stored at −80 °C. Cells were resuspended in 1 ml RLT buffer (Qiagen RNeasy Kit) and lysed by bead beating (2 × 1 min) at 4 °C followed by RNA extraction according to the manufacturer's instructions. cDNA was synthesized using Invitrogen Superscript III reverse transcriptase following the recommended protocol. Quantitative PCR reactions were performed in triplicate using Power SYBR Green PCR Master Mix (ABI) with either *C. difficile* 16 s (JP048–JP049)- or *treA* (CR045–CR046)-specific primers. Standard curves of cDNA were run to determine primer efficiencies and calculated as in ref. 27. Expression of *treA* was determined using an average of triplicate C_T values from each biological sample.

Mouse model of *C. difficile* infection. Humanized microbiota mice were derived from an initial population of germ-free C57BL/6 mice stably colonized with human gut microbiota and validated for use as a model of *C. difficile* infection[28]. Humanized microbiota mice aged 6–8 weeks of both sexes were treated with a five-antibiotic cocktail consisting of kanamycin (0.4 mg ml^{-1}), gentamicin (0.035 mg ml^{-1}), colistin (850 U ml^{-1}), metronidazole (0.215 mg ml^{-1}), and vancomycin (0.045 mg ml^{-1}) administered *ad libitum* in drinking water for 4 days. Water was switched to antibiotic-free sterile water and 24 h later mice were administered an intraperitoneal injection of clindamycin (10 mg per kg (body weight)). After a further 24 h, mice were challenged with 10^4 *C. difficile* spores by oral gavage. Sterile drinking water containing 5 mM trehalose was provided *ad libitum* (for the with or without trehalose study, mice were administered an additional 100 µl oral gavage of 300 mM trehalose daily) and mice were monitored for signs of disease.

In a separate experiment to determine *C. difficile* colonization load and toxin production, mice were euthanized 48 h after challenge with either R20291 or R20291Δ*treA*. *C. difficile* levels in caecal contents were determined by qPCR of toxin genes[10]. Relative toxin levels were assessed using a Vero Cell rounding assay[10]. Sample sizes for all experiments were determined using power analysis based upon previous experimental data. No randomization of animals was performed; however, all groups were checked to ensure no significant difference in the age, weight, or sex of mice between groups before starting experiments. All animal use was approved by the Animal Ethics Committee of Baylor College of Medicine (protocol number AN-6675). The investigators were not blinded to allocation during experiments and outcome assessment.

Detection of trehalose in caecal contents and human ileostomy fluid. Antibiotic-treated groups were pre-treated with the five-antibiotic cocktail for 3 days. Mice were gavaged with 100 µl of 5 mM trehalose, 300 mM trehalose, or water. Twenty minutes after gavage, mice were euthanized, and caecal contents harvested and vigorously mixed with two volumes/weight ice cold DMM (no carbohydrate). Supernatant was separated by centrifugation, filter sterilized, and reduced in an anaerobic chamber overnight before use. Ileostomy effluent from three anonymous donors was self-collected into sterile containers and stored at −20 °C until thawed, filter sterilized, and used for assay.

Strains were grown overnight and subcultured 1:50 into DMM supplemented with 20 mM succinate. Upon reaching an $A_{600\,\mathrm{nm}}$ of 0.2–0.3, cells were collected by centrifugation and resuspended in approximately 300 µl caecal or ileostomy fluid and incubated anaerobically for 30 min. Cells were then centrifuged and resuspended in RNALater (Invitrogen) before qRT–PCR analysis.

Bioreactor model for RT078 Δ*ptsT* competition. Faecal communities were established in continuous-flow minibioreactor arrays as previously described[10] using bioreactor defined medium[29] without starch (BDM4). Communities were disrupted by addition of clindamycin (250 µg ml^{-1}) continuously supplied in the medium for 4 days. After clindamycin treatment, communities were supplied BDM4 without clindamycin supplemented with trehalose (5 mM final concentration, $\mathrm{BDM4}_{\mathrm{tre}}$). After 1 day of growth in $\mathrm{BDM4}_{\mathrm{tre}}$, to allow washout of clindamycin, communities were challenged with a mixture of exponentially growing CD1015 strains (RT078 wild type and Δ*ptsT*). The competitive index was determined by dividing the proportion of wild-type cells at the end of the competition by the proportion at the start. The competitive index of wild type:Δ*ptsT* strains was determined by qPCR. The competitive index of wild type versus CD1015Δ*ptsT*::ahTCptsT was calculated by selective plating.

Isolation of spontaneous *treR* mutants. *C. difficile* strains were inoculated into continuous-flow minibioreactor arrays as previously described[10] using bioreactor defined medium[29] without starch (BDM4) supplemented with 5 mM trehalose ($\mathrm{BDM4}_{\mathrm{tre}}$). Every 24 h after the start of the experiment, 200 µl PBS containing 100 mM trehalose was spiked into each minibioreactor. The reactors were sampled daily, serially diluted, and plated to DMM agar supplemented with 10 mM trehalose. Resulting colonies were streak purified, and the ability to grow on low trehalose (10 mM) verified on plates and in broth culture. The *treR* gene was sequenced and compared with the isogenic parent strain.

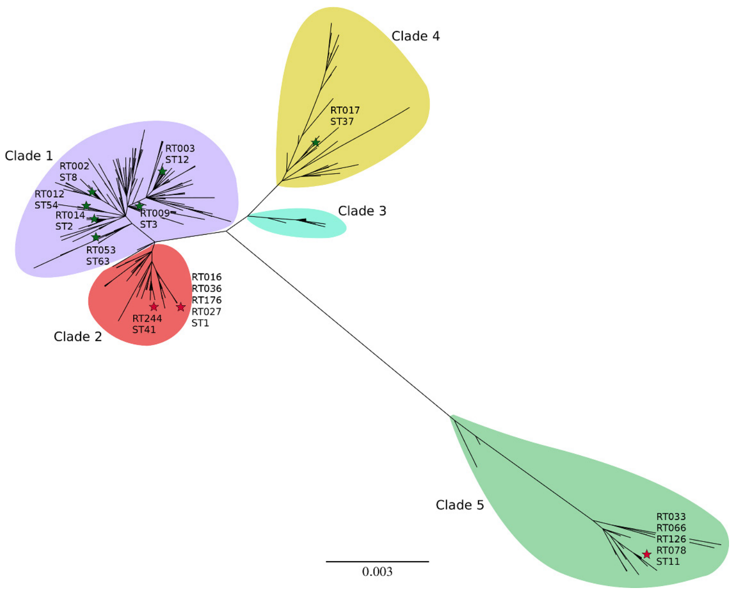
Statistics. Statistical analyses were performed using R (version 3.3.2). A Student's two-sample *t*-test (two-tailed) was used for comparisons of continuous variables between groups with similar variances; Welch's two-sample *t*-test (two-tailed) was used for comparisons of continuous variables between groups with dissimilar variances. *P* values from multiple comparisons were corrected using the Holm method[30]. A Wilcoxon rank-sum test with continuity correction was used for the toxin assay where data were non-normal. Fold-change data from *treA* gene expression experiments were log-normalized before statistical analysis. Data were visualized using individual data points and group means. Cox proportional hazards models and likelihood ratio tests were used to test significant differences in survival distributions among *C. difficile*-challenged groups of animals.

Collection of human bio-specimens. For faecal samples, live participants who were self-described as healthy and had not consumed antibiotics within the previous 2 months were recruited to provide faecal samples for human faecal bioreactor experiments. Informed consent was obtained before collection of samples and no identifying information was obtained along with the sample. Faecal samples were collected in sterile containers, transported to the laboratory on ice in the presence of anaerobic gas packs (BD Biosciences) within 16 h of collection, manually homogenized in an anaerobic environment, aliquoted into anaerobic tubes, and sealed and stored at −80 °C until use. Participants who had ileostomies placed owing to previous, undisclosed illnesses were recruited to provide ileostomy effluents. Informed consent was obtained before collection of samples and no identifying information was obtained with the sample. After transfer from the ostomy bag to a sterile collection container, ileostomy samples were transported to the laboratory on ice within 12 h of collection. Upon receipt, samples were stored at −20 °C. Ileostomy donors were recruited through the Ostomy Association of Greater Lansing, and were most probably residents of Lansing, Michigan, USA, and its surrounding counties. Samples were stored at −80 °C or −20 °C for 3–4 years before use. Samples were randomly selected for testing from a bank of available samples. Samples were collected according to a protocol approved by the Institutional Review Board of Michigan State University (protocol number 10-736SM).

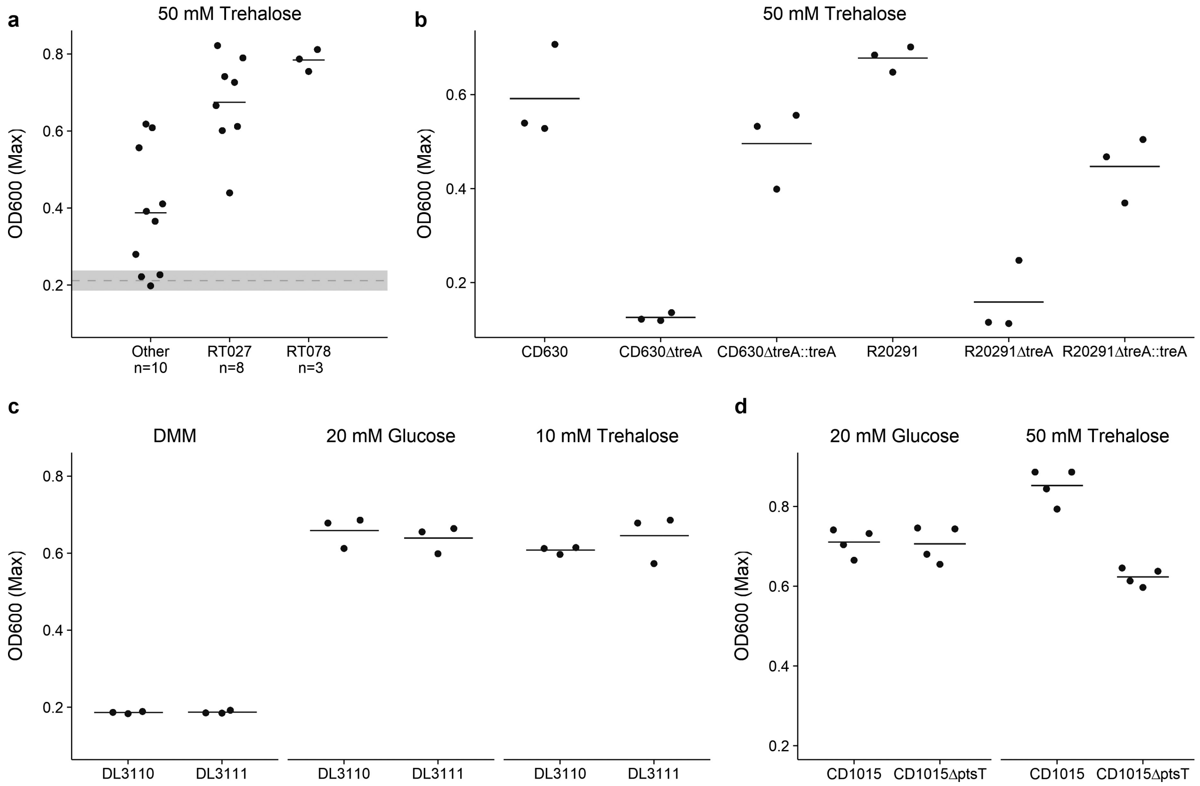
Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request. Source data for Figs 1–5 and Extended Data Figs 2, 3 and 5 are provided with the paper.

21. Theriot, C. M. *et al.* Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to *Clostridium difficile* infection. *Nat. Commun.* **5,** 3114 (2014).

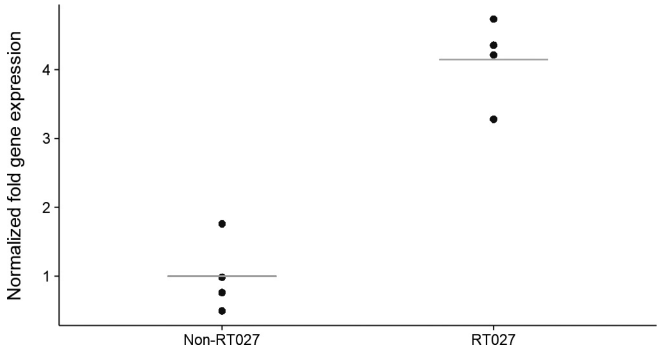
22. Knetsch, C. W. *et al.* Genetic markers for *Clostridium difficile* lineages linked to hypervirulence. *Microbiology* **157,** 3113–3123 (2011).
23. Bouillaut, L., Self, W. T. & Sonenshein, A. L. Proline-dependent regulation of *Clostridium difficile* Stickland metabolism. *J. Bacteriol.* **195,** 844–854 (2013).
24. Ng, Y. K. *et al.* Expanding the repertoire of gene tools for precise manipulation of the *Clostridium difficile* genome: allelic exchange using pyrE alleles. *PLoS ONE* **8,** e56051 (2013).
25. Fagan, R. P. & Fairweather, N. F. *Clostridium difficile* has two parallel and essential Sec secretion systems. *J. Biol. Chem.* **286,** 27483–27493 (2011).
26. de Kok, S. De *et al.* Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth. Biol.* **3,** 97–106 (2014).
27. Pfaffl, M. W. in *Real-time PCR* (ed. Dorak, T.) 63–82 (Taylor & Francis, 2006).
28. Collins, J., Auchtung, J. M., Schaefer, L., Eaton, K. A. & Britton, R. A. Humanized microbiota mice as a model of recurrent *Clostridium difficile* disease. *Microbiome* **3,** 35 (2015).
29. Auchtung, J. M., Robinson, C. D., Farrell, K. & Britton, R. A. in *Clostridium difficile: Methods and Protocols* (eds Roberts, A. P. & Mullany, P.) 235–258 (Springer, 2016).
30. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6,** 65–70 (1979).
31. Griffiths, D. *et al.* Multilocus sequence typing of *Clostridium difficile*. *J. Clin. Microbiol.* **48,** 770–778 (2010).
32. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33,** 1870–1874 (2016).
33. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
34. Roca, A. I., Abajian, A. C. & Vigerust, D. J. ProfileGrids solve the large alignment visualization problem: influenza hemagglutinin example. *F1000 Res.* **2,** 2 (2013).
35. Dingle, T. C., Mulvey, G. L. & Armstrong, G. D. Mutagenic analysis of the *Clostridium difficile* flagellar proteins, FliC and FliD, and their contribution to virulence in hamsters. *Infect. Immun.* **79,** 4061–4067 (2011).
36. Popoff, M. R., Rubin, E. J., Gill, D. M. & Boquet, P. Actin-specific ADP-ribosyltransferase produced by a *Clostridium difficile* strain. *Infect. Immun.* **56,** 2299–2306 (1988).
37. Smith, C. J., Markowitz, S. M. & Macrina, F. L. Transferable tetracycline resistance in *Clostridium difficile*. *Antimicrob. Agents Chemother.* **19,** 997–1003 (1981).



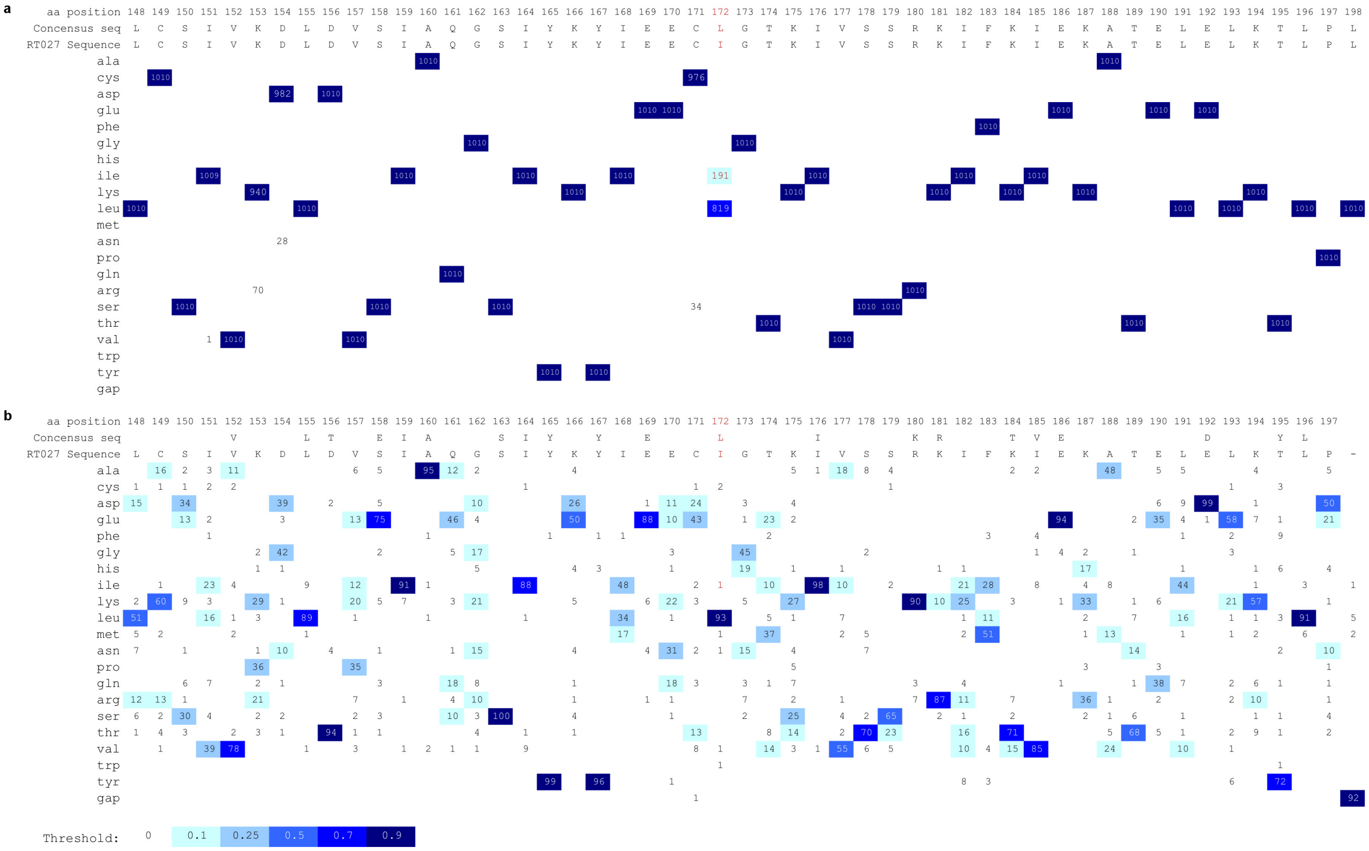
Extended Data Figure 1 | Phylogenetic organization of *C. difficile* MLST profiles. Maximum likelihood tree based upon concatenated multi locus sequence typing genes of the 399 current profiles available at https://pubmlst.org/cdifficile/[31]. Stars indicate position of strains used in this study, with red stars indicating sequence types possessing either the TreR L172I amino acid substitution (ST1, ST41) or four-gene insertion (ST11). Tree constructed using MEGA7 (ref. 32).



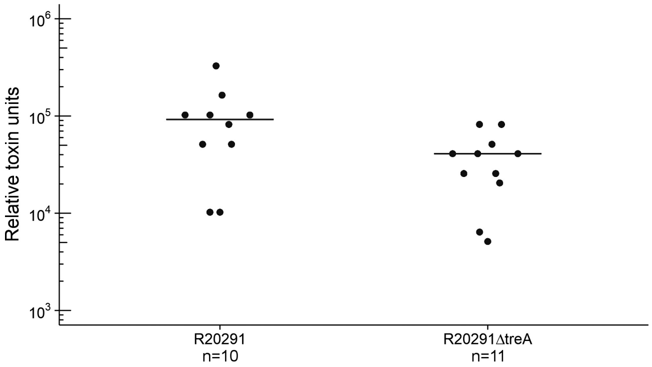
Extended Data Figure 2 | Growth of *C. difficile* strains. a, The majority of strains can grow on 50 mM trehalose. Dashed grey line and band indicate mean growth and s.d. in DMM without a carbon source. Solid lines indicate mean growth yield ($A_{600\,\mathrm{nm}}$) for groups: non-RT027/078 ($n = 10$), RT027 ($n = 8$), and RT078 ($n = 3$). **b**, Deletion of *treA* ablates the ability of both CD630 (RT12) and R20291 (RT027) to grow on trehalose. This phenotype can be restored by supplying *treA* on an inducible plasmid ($n = 3$ for each strain/group). **c**, RT244 strains (DL3110 and DL3111) possessing the TreR L172I mutation are capable of growth on 10 mM trehalose ($n = 3$ for each strain/group). **d**, CD1015Δ*ptsT* can metabolize 50 mM trehalose ($n = 4$ for each strain/group). For **a**–**d**, points represent biologically independent samples, solid bars are means.



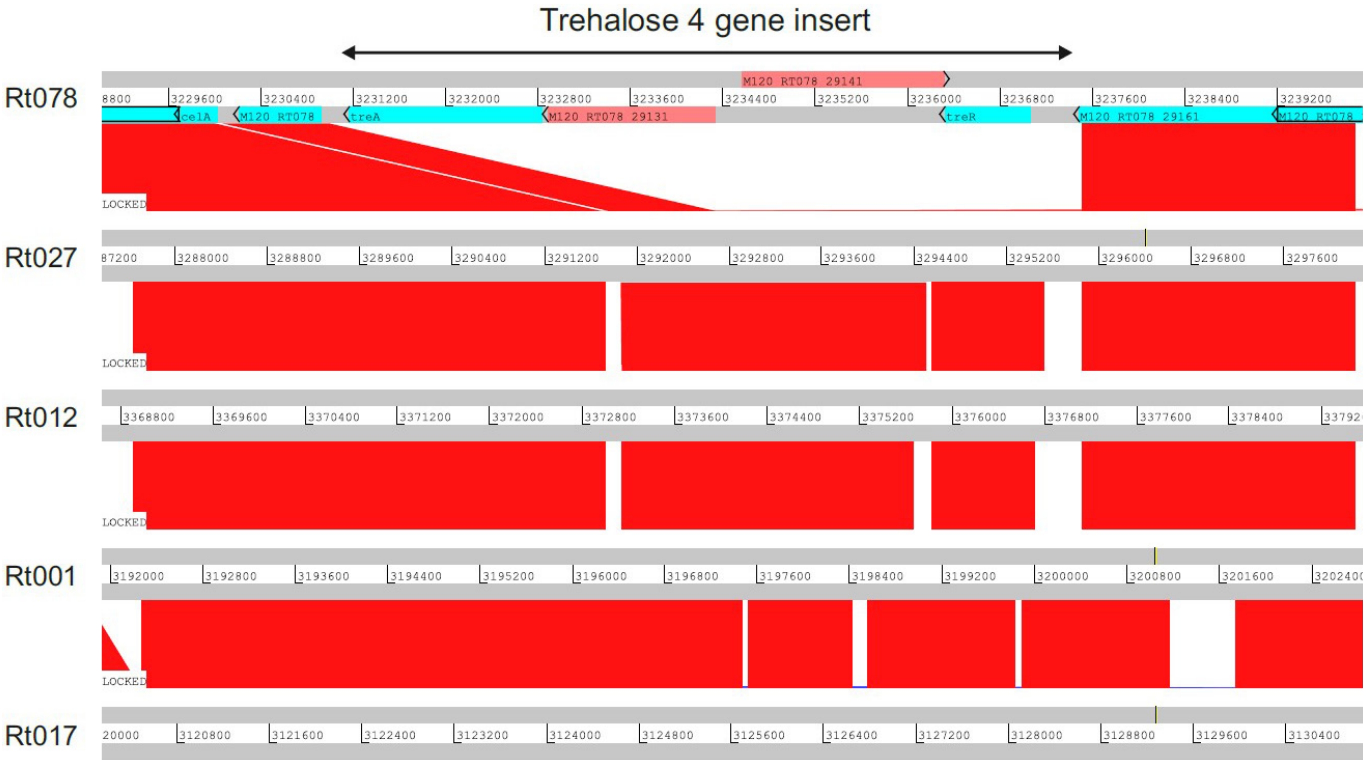
Extended Data Figure 3 | RT027 strains express *treA* at a significantly higher level than non-RT027 strains in the presence of 25 mM trehalose. Each data point ($n = 4$ ribotypes per group) represents gene expression from a different, biologically independent, strain and is an average from two to five independent experiments. $P = 0.029$, Mann–Whitney–Wilcoxon test (two-sided). Bar indicates mean expression.



Extended Data Figure 4 | RT027 strains have an L172I mutation at a highly conserved site. a, The *treR* genes from available *C. difficile* whole-genome sequencing files on the NCBI database (accessed 11 May 2017) were identified by tblastn and translated to protein sequences. Sequence fragments shorter than 240 amino acids were discarded and the remaining 1,010 sequences aligned with Clustal Omega[33]. All 191 sequences containing the L172I SNP also contained the *thyA* gene, a marker for the RT027 lineage; *thyA* was not found in any other genomes. Numbers indicate the number of sequences with a corresponding amino acid in that position. Multiple sequence alignment visualization generated with ProfileGrid[34]. **b**, The TreR protein sequence from RT027 strain R20291 was blasted against non-*C. difficile* sequences in the NCBI database and the top 99 matches (along with R20291$_{\text{treR}}$) aligned with Clustal Omega. The leucine at position 172 was found to be conserved in 93 of 99 non-*C. difficile* sequences. To confirm the importance of this residue, TreR was blasted against all non-clostridial sequences in the NCBI database and the top 500 hits saved. After removal of duplicate species, 191 sequences were aligned with Clustal Omega. The leucine at residue 172 was conserved in 83% of sequences (data not shown).

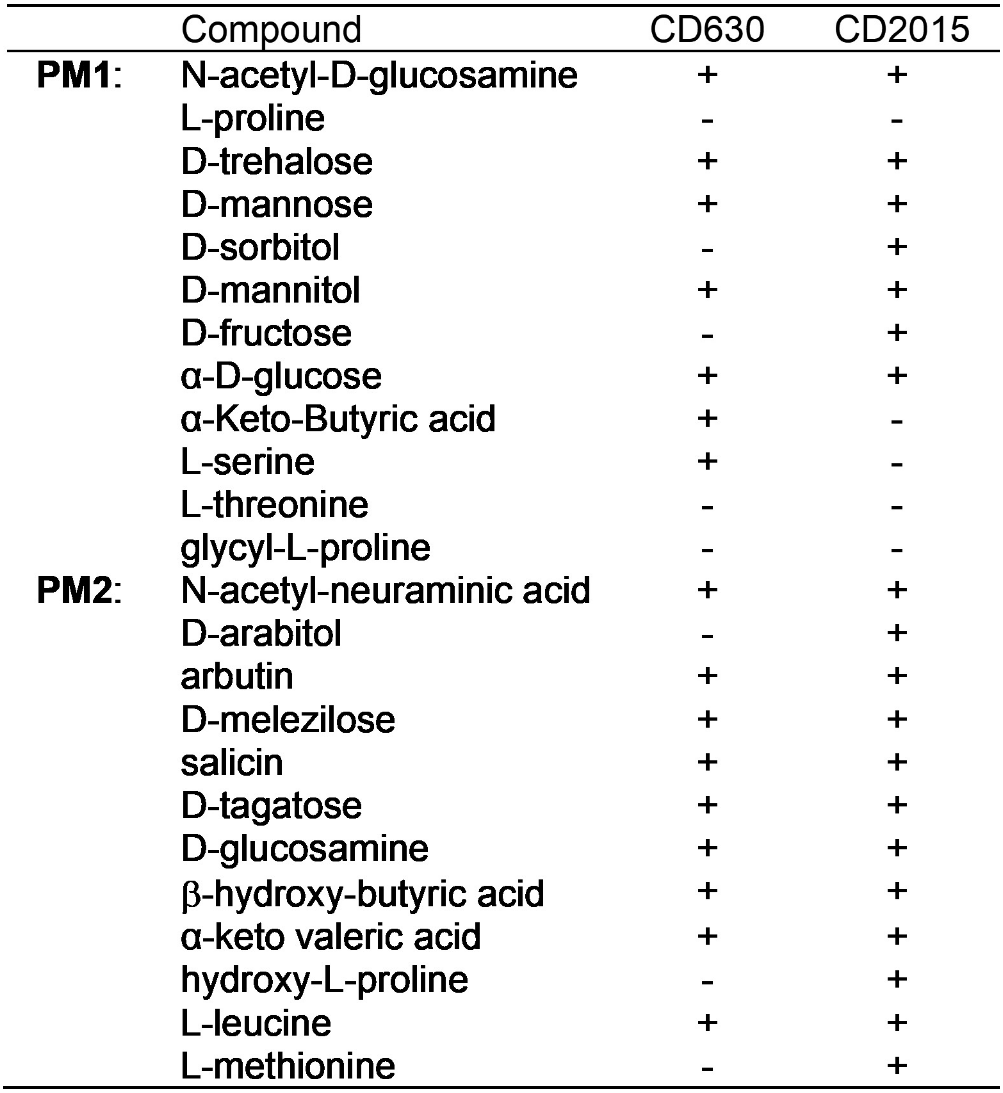


Extended Data Figure 5 | A *treA* knockout strain decreased toxin production 48 h after infection. Mice were gavaged with 10^4 spores of either R20291 or R20291Δ*treA* and provided with 5 mM trehalose in drinking water. Points represent toxin levels from individual mice (R20291 $n = 10$, R20291Δ*treA* $n = 11$) euthanized 48 h after infection. Bars are means. Mice gavaged with R20291Δ*treA* had significantly lower toxin levels ($P = 0.0268$; Wilcoxon–Mann–Whitney test (two-sided), median 40,960, IQR 23,040–46,080 versus 92,160, IQR 51,200–102,400).



Extended Data Figure 6 | The four-gene trehalose insertion is only present in the RT078 lineage. Artemis comparison tool displaying pairwise comparisons between *C. difficile* RT078 genome (M120) sequence and genome sequences from other *C. difficile* ribotypes (ribotypes indicated on the left). Numbers between grey bars indicate the genomic region where the trehalose four-gene insert is located (3231169–3237057). Regions of sequence homology are displayed in red. The trehalose four-gene insert of RT078 (indicated by the arrow on the top) was observed in RT078, but was absent in other ribotypes.

Extended Data Table 1 | Compounds conferring at least 1.5-fold growth advantage in Biolog Phenotypic Microarray plates PM1 or PM2



	Compound	CD630	CD2015
PM1:	N-acetyl-D-glucosamine	+	+
	L-proline	-	-
	D-trehalose	+	+
	D-mannose	+	+
	D-sorbitol	-	+
	D-mannitol	+	+
	D-fructose	-	+
	α-D-glucose	+	+
	α-Keto-Butyric acid	+	-
	L-serine	+	-
	L-threonine	-	-
	glycyl-L-proline	-	-
PM2:	N-acetyl-neuraminic acid	+	+
	D-arabitol	-	+
	arbutin	+	+
	D-melezilose	+	+
	salicin	+	+
	D-tagatose	+	+
	D-glucosamine	+	+
	β-hydroxy-butyric acid	+	+
	α-keto valeric acid	+	+
	hydroxy-L-proline	-	+
	L-leucine	+	+
	L-methionine	-	+

Results of individual experiments where growth was (+) or was not (–) increased by at least 1.5-fold over DMM control.

Extended Data Table 2 | Spontaneous *C. difficile* mutants able to utilize 10 mM trehalose

Strain	Ribotype	Nonsense mutation*	Missense mutation	Insertions/ deletions	Number of independent isolates
3014	001	18	-	-	1
2012	002	63	-	-	1
2012	002	89	-	-	2
2012	002	15	-	-	1
2012	002	22	-	-	1
2012	002	-	S20I	-	1
2012	002	64	-	-	1
2012	002	24	-	-	1
2012	002	20	-	-	1
1014	014	-	S41I, T118K	1	1
1014	014	-	T118K	-	1
2048	053	70	-	-	1

*Numbers refer to the positions in the consensus TreR amino acid sequence that become a premature stop codon.

Extended Data Table 3 | Strains, primers and plasmids

Strains	Ribotype	MLST (Clade)	Note	Reference
CD630	12	54 (1)	Erythromycin sensitive	35
CD630Δ*treA*	12	54 (1)		This Study
CD630 pRFP185-P_{aTC}-*ptsT*	12	54 (1)		This Study
CD630Δ*treA* pRFP185-P_{aTC}-*treA*	12	54 (1)		This Study
R20291	27	1 (2)		24
R20291Δ*pyrE*	27	1 (2)		24
R20291Δ*treA*	27	1 (2)		This Study
R20291Δ*treA* pRFP185-P_{aTC}-*treA*	27	1 (2)		This Study
CD1015	78	11 (5)		†
CD1015Δ*pyrE*	78	11 (5)		This Study
CD1015Δ*ptsT*	78	11 (5)		This Study
CD1015Δ*ptsT* pRFP185-P_{aTC}-*ptsT*	78	11 (5)		This Study
VPI10463	3	12 (1)	High toxin producer	
CD196	27	1 (2)	Ancestral RT027 strain	36
CD1007	053-163	63 (1)		†
CD1014	014-20	2 (1)		10
CD2012	2	8 (1)		†
CD2018	unique UM isolate			†
CD2046	unique UM isolate			†
CD2048	053-163	63 (1)		10
CD37	9	3 (1)	Non-toxigenic strain	37
CD4004	2	8 (1)		†
CD4011	1	3 (1)		†
CD2015	27	1 (2)		10
CD3017	27	1 (2)		10
CD4010	27	1 (2)		10
CD4012	27	1 (2)		†
CD4015	27	1 (2)		10
CD2001	78	11 (5)		†
CD2058	78	11 (5)		†
DL3110	244	41 (2)	contains L172I in treR	11
DL3111	244	41 (2)	contains L172I in treR	11

Primers	Note		Sequence 5′ - 3′
IBS1.2	Insertional deletion of *treA* in CD630		atatcaagcttttgcaacccacgtcgatcgtgaatagaagattattgtgcgcccagatagggtg
EBS1			cagattgtacaaatgtggtgataacagataagtcattattattaacttacctttctttgt
EBS2			cgcaagtttctaatttcggttttctatcgatagaggaaagtgtct
CR064	CD630 *treA* insertion check		gcaacaatgatggtataggtgatataaatgg
CR065			ggaacagaaccatcaggtttagca
JCb092	Upstream homology arm R20291 *treA*	5′ phosphorylated	agacctttaaggagggataggggt
JCb093		5′ phosphorylated	aggagaaacgtacataggagtcaacca
JCb094	Downstream homology arm R20291 *treA*	5′ phosphorylated	cctgaaacttatttgaataaattaaactacac
JCb095		5′ phosphorylated	gtttgatactgatggagggcctta
JCb096	Bridging oligos for ligase cycling		gccctccatcagtatcaaacggggatcctctagagtcgac
JCb097			ctcctatgtacgtttctcctcctgaaacttatttgaataa
JCb098			cgaattcgagctcggtacccagacctttaaggagggatag
JCb135	Upstream homology arm CD1015 *pyrE*	*SbfI*	atacctgcaggagggacatttttattatcttcag
JCb136		5′ phosphorylated	acaacatcttcagcaattattatctttg
JCb137	Spacer from pMTL-YN2	5′ phosphorylated	gcggccgctgtatccatatgacc
JCb138		5′ phosphorylated	actagcgccattcgccattcagg
JCb139	Downstream homology arm CD1015 *pyrE*	5′ phosphorylated	gcggccgctgtatccatatgacc
JCb140		*AscI*	atatggcgcgccataacattaataaaatttaaaatcaataattat
JCb141	Bridging oligos for ligase cycling		ataattgctgaagatgttgtgcggccgctgtatccatatg
JCb142			gaatggcgaatggcgctagttaataaaaacttaattattt
JCb153	Upstream homology arm CD1015 *ptsT*	*EcoRI*	atagaattcaaggacccaggaatttgacc
JCb154		*BamHI*	ataggatccatctacttatcctttctctttttttattataag
JCb155	Downstream homology arm CD1015 *ptsT*	*BamHI*	ataggatcccaaatgacaatatataaatataattcccttgg
JCb156		*NcoI*	ataccatggcgtggttggtcatggttaca
JCb167	CD1015Δ*ptsT* check		cggaatttcttttatattcatttgg
JCb168			cccaatttgttggagcactt
JCb225	Conformation of *pyrE* knockout/repair		atgggaatgggcggaataac
JCb226			gcttggaagcagctacaacaga
JCb211	CD1015 *pyrE* repair	*NotI*	atagcggccgcttacattcctaattccttgaactctc
JP048	qPCR for *C. difficile* 16S DNA		ttgagcgatttacttcggtaaaga
JP049			ccatcctgtactggctcacct
CR045	qPCR for *C. difficile treA* DNA		tacgctgatggtcctcgtat
CR046			cgcctcctttataatctgttttc

Plasmids	Reference
pMTL-YN2	24
pMTL-YN2C	24
pMTL-YN4	24
pRPF185	25
Plasmid Maps	
pJC-R20291*treA*KO	https://benchling.com/s/seq-DANCiRRu7FwNqRJCO9iu
pJC-CD1015*pyrE*KO	https://benchling.com/s/seq-Km3QbgAkU66DkNtaOA4R
pRFP185-P_{aTC}-*ptsT*	https://benchling.com/s/seq-hKTbvE2pfE8MpvvEqFce
pJC-CD1015*pyrE*Repair	https://benchling.com/s/seq-d25nmFmvSPtR1iQB8vka
pRFP185-P_{aTC}-*treA*	https://benchling.com/s/seq-TTthHzlU2fsfRZ94oD9V

†Clinical isolates obtained from the Michigan Department of Community Health. Collected from Michigan hospitals between December 2007 and May 2008. (References 10, 11, 24, 25, 35–37 are cited in the table.)

doi:10.1038/nature25440

Molecular mechanism of promoter opening by RNA polymerase III

Matthias K. Vorländer[1], Heena Khatter[1], Rene Wetzel[1], Wim J. H. Hagen[1] & Christoph W. Müller[1]

RNA polymerase III (Pol III) and transcription factor IIIB (TFIIIB) assemble together on different promoter types to initiate the transcription of small, structured RNAs. Here we present structures of Pol III preinitiation complexes, comprising the 17-subunit Pol III and the heterotrimeric transcription factor TFIIIB, bound to a natural promoter in different functional states. Electron cryo-microscopy reconstructions, varying from 3.7 Å to 5.5 Å resolution, include two early intermediates in which the DNA duplex is closed, an open DNA complex, and an initially transcribing complex with RNA in the active site. Our structures reveal an extremely tight, multivalent interaction between TFIIIB and promoter DNA, and explain how TFIIIB recruits Pol III. Together, TFIIIB and Pol III subunit C37 activate the intrinsic transcription factor-like activity of the Pol III-specific heterotrimer to initiate the melting of double-stranded DNA, in a mechanism similar to that of the Pol II system.

The eukaryotic genome is transcribed by three nuclear DNA-dependent RNA polymerases, with Pol III transcribing short, structured RNAs including tRNAs, 5S rRNA and spliceosomal U6 RNA. Large numbers of Pol III transcripts are necessary to enable protein synthesis during cell growth and division, but Pol III must also be carefully regulated[1], as increases in its transcriptional activity have been associated with malignant transformation[2–6].

The principal transcription initiation factor of Pol III is TFIIIB, a complex consisting of three subunits: TATA-binding protein (TBP), B-related factor 1 (Brf1) and B double prime 1 (Bdp1). TFIIIB is positioned to bind to DNA either by a strong upstream TATA box, as in the case of the yeast type III genes, or by the six-subunit assembly factor TFIIIC. TFIIIB bound to promoter DNA is capable of directing several rounds of transcription *in vitro*[7–9].

TFIIIB is conserved from yeast to humans, and components of TFIIIB show homology with the general transcription factors of other RNA polymerases[10]. TBP is shared between the Pol I, Pol II and Pol III transcription machineries, although it is not essential for Pol I *in vitro*[11]. The N-terminal half of Brf1 is homologous with the Pol II general transcription factor TFIIB and the Pol I TFIIB-related factor Rrn7, whereas the C-terminal half of Brf1 is the major interaction hub that holds together the trimeric TBP–Brf1–Bdp1 complex[12]. Sequence analysis has identified three conserved blocks in the C terminus[13]: homology domain I (residues 286–304), homology domain II (residues 439–515) and homology domain III (residues 570–596) (Fig. 1a).

Bdp1 contains a highly conserved SANT (Swi3, Ada2, N-Cor and TFIIIB) domain and a DNA-binding linker[14], but is predicted to be otherwise disordered. However, three essential regions (ERs) have been defined[15] (Fig. 1a). The C-terminal ER I (residues 372–487) contains the SANT domain and flanking regions. ER II (residues 269–312) cross-links to promoter DNA[16] and is required for TFIIIC-dependent transcription *in vitro*[17]. Footprinting and hydroxyl-radical probing experiments suggested that ER I and ER II are buried in the preinitiation complex (PIC), whereas the N-terminal ER III (residues 158–252) is more accessible[15,17].

Pol III comprises 17 subunits that include the conserved 10-subunit core, the stalk and two additional subcomplexes, namely the C53–C37 heterodimer and the C82–C34–C31 heterotrimer. The heterodimer is located at the Pol III lobe and is considered to be homologous with the Pol II general transcription factor TFIIF and the Pol I subcomplex A49–A34.5. It has been demonstrated to be essential for the accurate initiation and termination of transcription[18–20].

The C82–C34–C31 heterotrimer sits on top of the clamp head and is considered to be homologous with the Pol II general transcription factor TFIIE. C34 contains three winged-helix (WH) domains, but the first two domains are disordered in the available electron cryo-microscopy (cryo-EM) structures[21]. C34 has been shown to be important for the interaction with TFIIIB and for open complex (OC) formation[22]. In order to understand how the Pol III-specific subunits interact and facilitate the initiation of transcription, it is necessary to study them in the context of the PIC containing TFIIIB.

Here we report the cryo-EM reconstruction of *Saccharomyces cerevisiae* Pol III–TFIIIB complexes in different functional states. The cryo-EM maps range from 3.7 Å to 5.5 Å resolution (Extended Data Fig. 1), and comprise the PIC bound to a DNA:RNA scaffold (initially transcribing complex, ITC), a spontaneously OC, and two reconstructions of a closed DNA complex (CC). Our results explain how Pol III-specific core subunits and TFIIIB engage in an intricate interaction network to recognize, bind and stabilize upstream promoter DNA, and which structural rearrangements occur during the CC-to-OC transition.

Cryo-EM of Pol III preinitiation complexes

Pol III PICs were assembled with recombinant Bdp1, Brf1–TBP fusion protein[23] and 81-nucleotide (nt) DNA scaffolds containing the U6 promoter, and were cross-linked with glutaraldehyde. Scaffolds were either fully complementary (CC) or contained a mismatch from base pair (bp) −7 to +8 relative to the transcription start site and a 6-nt RNA oligonucleotide complementary to +1 to +6 (ITC scaffold, Fig. 1a). After particle classification (Extended Data Fig. 2), the ITC scaffold yielded a reconstruction at 4.3 Å resolution with well-defined DNA and RNA visible at a lower threshold, suggesting partial occupancy owing to dissociation during sample preparation or cleavage by C11.

The CC dataset gave rise to reconstructions of a spontaneously formed OC and two CC states. The OC was determined at 3.7 Å resolution and is almost identical to the ITC, but shows only fragmented density for the template strand in the active site. The two CC reconstructions (CC1 and CC2, at 5.5 Å and 4.2 Å resolution, respectively) differ

[1] European Molecular Biology Laboratory (EMBL), Structural and Computational Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

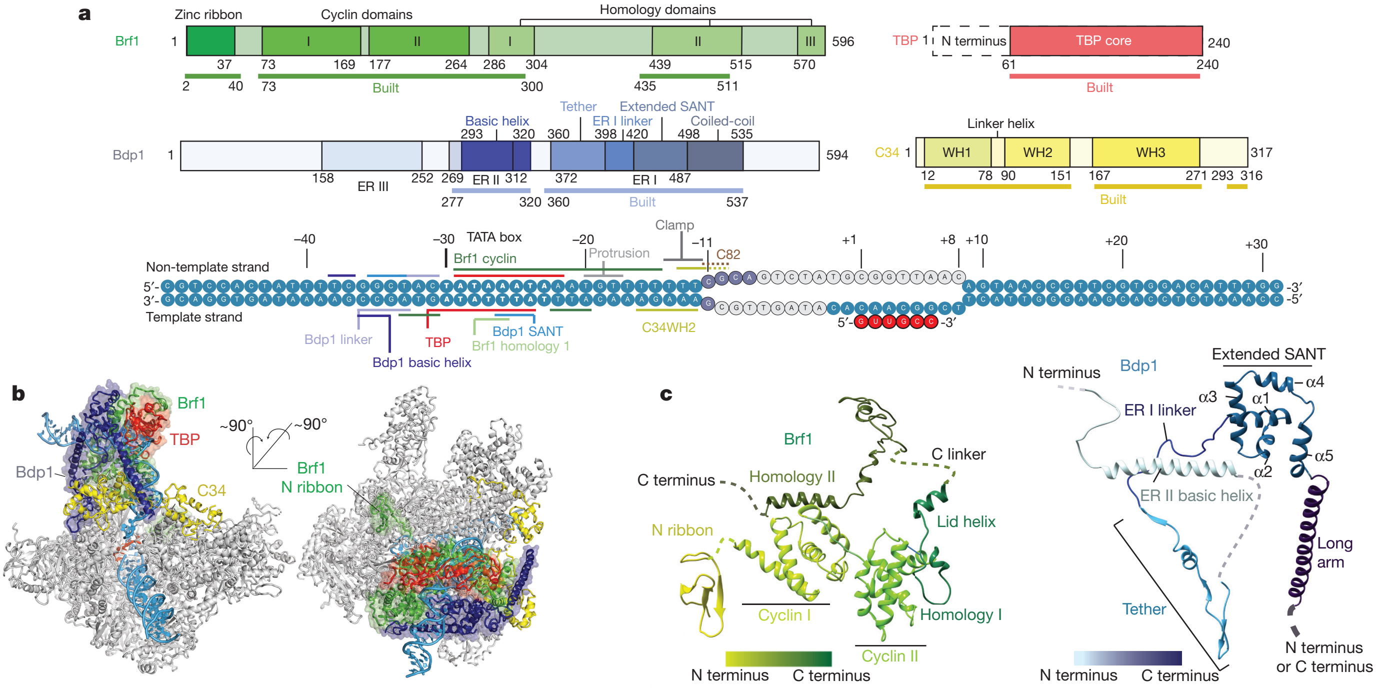


Figure 1 | Cryo-EM structure of the Pol III PIC. a, Domain organization and conserved regions in TFIIIB and Pol III subunit C34 (top), and the DNA transcription scaffold used in this study (bottom). The sections of TFIIIB and C34 that are included in the ITC model are underlined and marked as 'built'. The contacts of TFIIIB and Pol III with the DNA transcription scaffold are indicated. Disordered DNA bases are depicted in light grey. **b**, Two views of the ITC model. TFIIIB is represented in ribbon form and as a transparent surface, C34 is shown in yellow, and other Pol III subunits are coloured grey. **c**, Ribbon representation of the structures of Brf1 (top) and Bdp1 (bottom).

in the orientation of TFIIIB with respect to Pol III, and show a well-defined Pol III core with markedly lower resolution at the periphery (Extended Data Fig. 3).

We built atomic models of the PIC in the CC, OC and ITC states. Local amplitude scaling (LocScale)[24] of our maps supported model building. The TFIIIB structure comprises most of Brf1 (Fig. 1a), whereas for Bdp1 it includes ER II (residues 275–320) and ER I (residues 365–537). In addition, the WH1 and WH2 domains of C34, and a previously disordered 'initiation/termination loop' (residues 211–224) in subunit C37, which has been shown to have important roles in both transcription initiation and termination[19,20,25,26], become ordered in the ITC and OC structures.

Structure of TFIIIB in the PIC

TFIIIB adopts an intricate fold in the PIC (Figs 1c, 2). The N terminus of Brf1 forms a zinc ribbon that interacts with the Pol III dock (Fig. 1b),

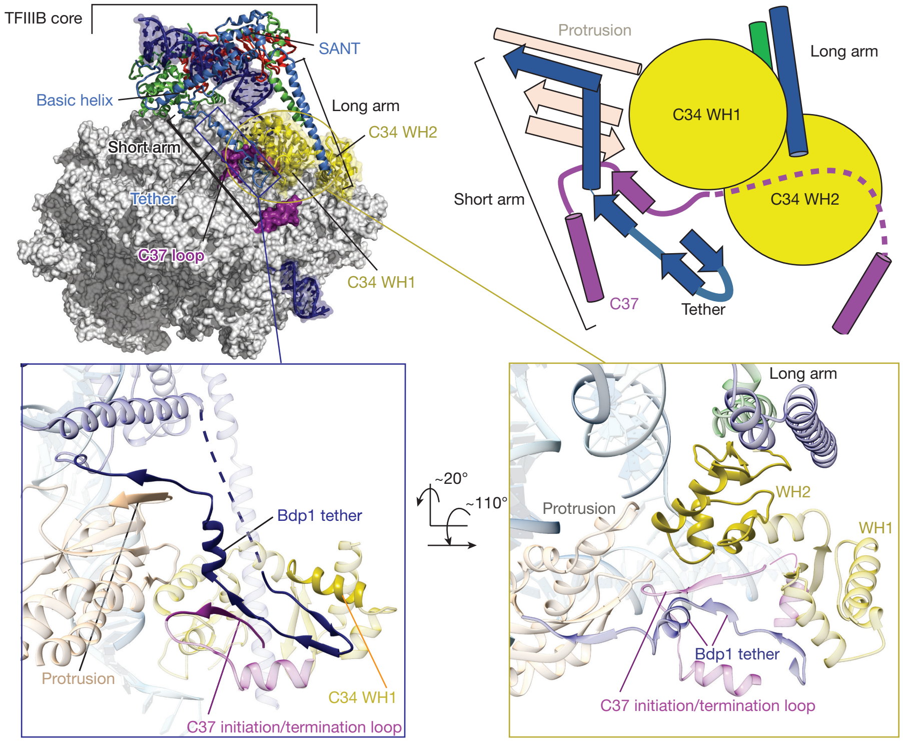


Figure 2 | Interactions of TFIIIB with C34 and C37 in the PIC. Top, location of TFIIIB on Pol III (left) and schematic view of interactions that stabilize C34 in the position observed in the OC and ITC structures (right). Bottom, close-up view of the Bdp1 tether (left), with interactions between the Bdp1 tether, Pol III protrusion, C37 initiation/termination loop, and C34 WH1 and WH2 highlighted; regions that contact the Bdp1 tether are depicted in solid colour, other elements are transparent. Close up view of C34 WH2 (right); C34 WH2 is stabilized by both arms of TFIIIB, Pol III protrusion and C37.

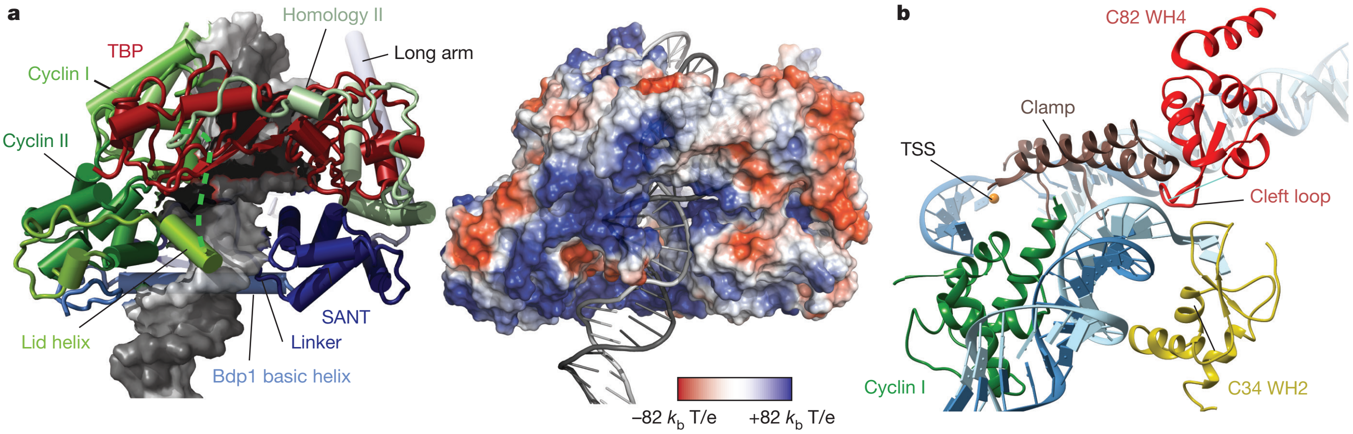


Figure 3 | Protein–DNA contacts in the Pol III PIC. a, TFIIIB encircles promoter DNA. Cylinder representation of TFIIIB around the TATA box with DNA represented as the surface (left), and electrostatic potential mapped onto the TFIIIB structure (right). **b**, The transcription bubble is stabilized by Brf1 cyclin I, the Pol III clamp, C34 WH2 and C82 WH4. The non-template DNA strand is stabilized by the C82 cleft loop and C34 WH2. TSS, transcription start site.

and the cyclin domains I and II bind the Pol III wall and upstream promoter DNA (Fig. 1). The Brf1 homology domain I, which follows cyclin II, forms a short 'lid' helix that lies on top of the DNA (Figs 1c, 3a). The C linker connects homology domains I and II but exhibits only weak density in proximity to the Bdp1 SANT domain, in agreement with photo-crosslinking and mutagenesis studies[27].

Brf1 homology domain II is in the same conformation as described in the crystal structure of TBP and Brf1 (residues 437–506)[28] (Fig. 3a). We could not model any residues after homology domain II, although we observe weak density close to C34 WH3 (Extended Data Fig. 4); this may correspond to Brf1, because Brf1 residue 549 photo-crosslinks to C34 (ref. 27).

The most N-terminal, ordered part of Bdp1 corresponds to ER II, in which residues 277–292 interact with cyclin II of Brf1, and residues 293–319 form a 'basic' helix containing 14 positively charged residues that runs behind upstream DNA (Figs 1c, 3). The residues connecting ER II to ER I are disordered. ER I is located between C34 WH2 and the heterodimer, forming a β-hairpin that interacts with WH1 and WH2 of C34, and with the C37 initiation/termination loop (Fig. 2), in agreement with photo-crosslinking data[25,26]. Bdp1 then forms a short helix that runs along the protrusion domain of Pol III, followed by a β-strand that extends the anti-parallel β-sheet of the Pol III protrusion by one parallel strand. This section of Bdp1 (residues 360–398)—comprising the β-hairpin, the helix and the β-strand—interacts with three distinct Pol III elements: the protrusion, the previously disordered WH1 and WH2 of C34, and the C37 initiation/termination loop (Fig. 2). We name this element the Bdp1 tether, owing to its central position in the PIC that bridges between distant and/or mobile Pol III subunits.

The ER I linker, which is enriched in aromatic residues, inserts into the minor groove of the DNA and connects the tether with the SANT domain (Extended Data Fig. 4). The canonical SANT domain fold is extended by two C-terminal helices (Figs 1, 3a) leading into a long helix (residues 498–536) that forms a coiled-coil with Brf1 homology domain II and ends next to C34 WH2 and WH3.

In summary, TFIIIB forms a compact core centred around the TATA box, from which two arms emerge. A short arm, formed by the Bdp1 tether, and a long arm, formed by the Bdp1–Brf1 coiled-coil, bridge the TFIIIB core with the Pol III heterotrimer and heterodimer.

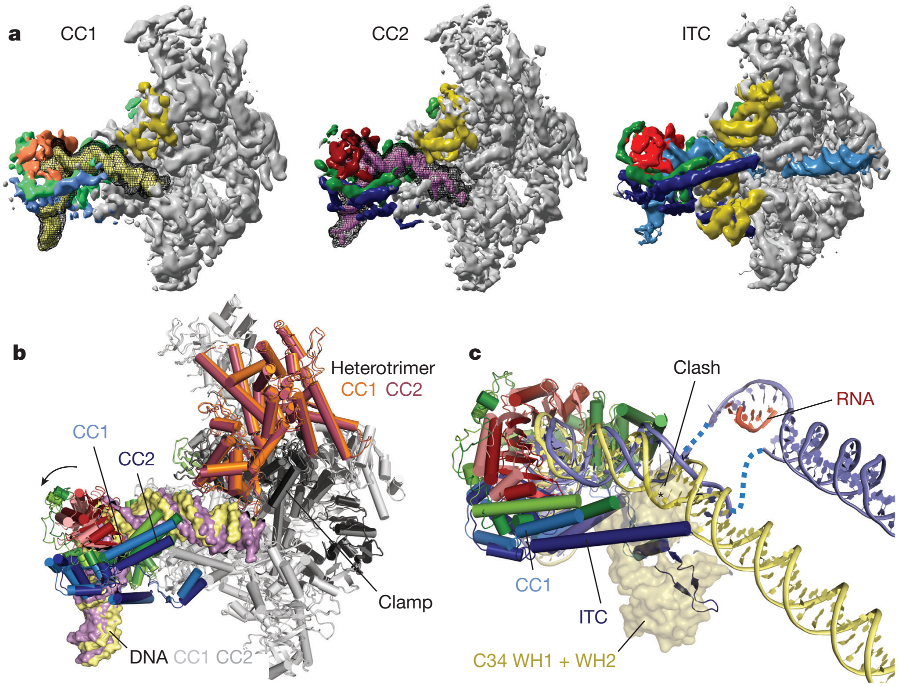


Figure 4 | Comparison between Pol III CC and OC structures. a, Electron microscopy densities of the CC1, CC2 and ITC, with densities corresponding to TBP coloured in orange or red, Brf1 green, Bdp1 blue and C34 yellow. CC1 is coloured in lighter shades and DNA is coloured in pale yellow for CC1 and magenta for CC2. DNA density before B-factor sharpening is shown as black mesh for CC1 and CC2. In the CCs, C34 WH1 and WH2 and the C-terminal section of the Bdp1 coiled-coil helix are disordered. **b**, Superposition of CC1 and CC2 models shows movement of TFIIIB and the heterotrimer, closing the cleft slightly during the CC1-to-CC2 transition. **c**, Superposition of the CC1 and ITC models. DNA in CC1 was further extended with B-DNA to assist with visualization. The positions of C34 WH1 and WH2 (transparent surface) in the ITC would clash with CC DNA, as indicated with an asterisk. All models were superimposed on C160.

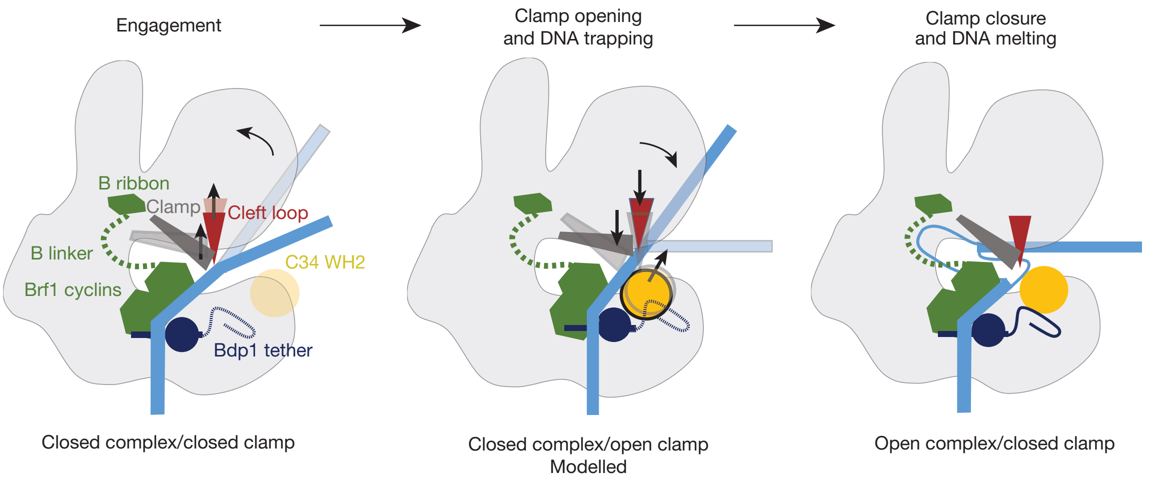

Engagement
Clamp opening and DNA trapping
Clamp closure and DNA melting
B ribbon
Clamp
Cleft loop
B linker
C34 WH2
Brf1 cyclins
Bdp1 tether
Closed complex/closed clamp
Closed complex/open clamp
Modelled
Open complex/closed clamp

Figure 5 | Mechanism of promoter opening and DNA melting in Pol III. Schematic of the promoter opening. Initially, the closed DNA is bent away by the C82 cleft loop and the clamp, while the Bdp1 tether and C34 WH1 and WH2 are disordered. Opening of the clamp enables the closed DNA to slide between clamp and lobe. Subsequently, C34 WH1 and WH2 become ordered and enclose double-stranded DNA. Closing of the clamp leads to the downwards movement of the cleft loop and DNA melting, while establishing stabilizing interactions between the Bdp1 tether and C34 WH1 and WH2. The non-template strand is stabilized by Pol III subunits C34 and C82; the template strand is presumably stabilized by the Brf1 linker.

In the TFIIIB structure, Bdp1 appears to compete with TFIIIC for the same binding sites on Brf1 (ref. 29) (Extended Data Fig. 5), rationalizing both the requirement of Bdp1 ER II for TFIIIC-dependent transcription[17] and the conformational changes of TFIIIC after the incorporation of Bdp1 (ref. 30).

Our structure reveals DNA-binding elements absent from previous crystallographic studies of TFIIIB[14,28] and Pol II paralogues, namely the Brf1 lid helix and the Bdp1 basic helix. Together with the Brf1 cyclin folds, TBP and the Bdp1 SANT and linker domains, they form a positively charged ring around the TATA box (Fig. 3), explaining why promoter bound TFIIIB is so unusually stable[8,17,31,32] and can serve as a 'roadblock' for Pol II transcription and DNA replication by DNA polymerase[33]. This 'headlock' arrangement illustrates how Bdp1 causes the kinetic trapping of the TFIIIB–DNA complex[34]. Accordingly, the term 'proteinaceous cage', which was coined to describe the TFIIIB–DNA structure almost two decades ago[34], is an excellent description of this structure.

TFIIIB coordinates promoter opening

TFIIIB triggers conformational changes in Pol III that are specific to the preinitiation state and/or render Pol III ready for elongation. In the PIC, Pol III adopts a conformation similar to that of the 'closed clamp'[21] state, although the heterotrimer shifts towards upstream DNA, slightly closing the clamp, and the stalk moves towards the heterotrimer, possibly mediated by the proposed C31–stalk contact[21] (Extended Data Fig. 6). The C53–C37 heterodimer contributes to the stabilization of the PIC by (partial) ordering of the initiation/termination loop, which interacts with the Bdp1 tether and inserts between WH1 and WH2 of C34, concomitant with the extension of the adjacent C37 helix (residues 197–202). The C-terminal domain of C11 is disordered, resembling the elongating state. In the active site, the rudder and the trigger loop are disordered and the bridge helix is bent, as in elongating Pol III (Extended Data Fig. 7a).

In the ITC, DNA was melted after position −11 relative to the transcription start site, and we modelled four nucleotides of the unpaired non-template strand that were opened in our preparation (Figs 1a, 3b). The upstream bubble edge and adjacent duplex DNA (bp −20 to −10) are stabilized by four different elements: C34 WH2, C82 WH4 including the cleft loop, the clamp core of C160, and Brf1 cyclin I (Fig. 3b). Notably, most of these elements are Pol III subunits, with little contribution from TFIIIB.

Our structure explains how, in addition to Pol III recruitment, DNA-bound TFIIIB also facilitates promoter opening. TFIIIB lacking the Brf1 C-terminal moiety (Δ283–596) is transcriptionally active[32] but highly unstable. This agrees with our findings that the C-terminal half of Brf1 (and Bdp1) has a scaffolding function that stabilizes the C34 winged-helix domains; however, it is not essential for DNA melting.

Brf1 and Bdp1 contribute to DNA opening in distinct ways, as has been probed by using mismatched DNA templates that rescue the phenotype of TFIIIB mutants[35]. Brf1 functions by extending the initial transcription bubble, which is thought to nucleate around bp −9 (ref. 36) and propagate towards the transcription start site; a Brf1 Δ1–68 mutant, lacking the zinc ribbon and N linker, can be rescued by introducing a mismatch between bp +2 and +6 or between −4 and +1 (ref. 35). These regions are close to the beginning of the N linker, which is disordered in our structure but has been shown previously to interact with the template strand in TFIIB[37,38].

The Bdp1 tether is required for strand separation at the upstream edge of the transcription bubble, as Bdp1 Δ355–372 can be rescued by introducing a mismatch in the DNA between bp −9 and −5.

The spatial organization of the PIC, with the Brf1 zinc ribbon and N linker close to the active site, and the Bdp1 tether providing a C34 binding platform close to the emerging transcription bubble, explains how the TFIIIB coordinates the opening of the promoter.

Closed DNA complexes

The Pol III closed DNA structures (Fig. 4) show early engagement intermediates of the PIC, as several of the stabilizing interactions seen in the OC and ITC are not yet established. The C34 WH1 and WH2 are disordered, as are the Bdp1 tether and most of the long arm of TFIIIB and the C37 initiation/termination loop. This suggests that C34 and the TFIIIB arms mutually stabilize each other, and that these interactions can occur only once the DNA has been melted or moved to slightly enter the cleft (discussed later). Accordingly, the resolution in our CC maps declines more strongly in the peripheral regions around the closed DNA and the heterotrimer compared to the OC and ITC maps (Extended Data Fig. 3).

The upstream DNA is kinked away by the clamp head and the C82 cleft loop in the CCs, resulting in a 30° bend introduced around bp −15. The DNA projects away from Pol III, and no density is visible after bp −2 (bp −5 in CC2), presumably because of the lack of stabilizing protein–DNA contacts (Fig. 4).

The CC1 and CC2 reconstructions differ in the orientations of TFIIIB, upstream DNA and the heterotrimer. In CC1, the heterotrimer is moved away from upstream DNA compared to its orientation in CC2, which slightly opens the clamp (Fig. 4b). In addition, TFIIIB is shifted by up to 6 Å, which moves the DNA away from the cleft (Fig. 4c), whereas CC2 exhibits an overall conformation similar to that in the OC and ITC states. This suggests that CC2 represents a later stage in the initiation process. The cryo-EM density of DNA extends further in the CC1 map, whereas that of TFIIIB is of better quality in the CC2 map.

The heterotrimer is TFIIF- and TFIIE-like

We compared the structure of the Pol III OC to that of the yeast Pol II OC (RCSB Protein Data Bank (PDB): 5FYW), in order to better understand the roles of general transcription factors and transcription factor-like subcomplexes in both systems (Extended Data Fig. 8). The overall topology of the PICs is highly similar in Pol III and Pol II, as both use an upstream promoter assembly centred on the TATA box that introduces a 90° kink, and a downstream assembly in which several winged-helix domains position the upstream bubble edge along the cleft. In the Pol III PIC, the upstream assembly consists of the TFIIIB core, and in the Pol II PIC it comprises TFIIB, TBP (or the much larger TFIID) and TFIIA.

The downstream assembly in Pol III is provided by the heterotrimer, in which the transcription bubble is stabilized by the C34 WH2 and the C82 cleft loop (Fig. 3b, Extended Data Fig. 8b). In the Pol II PIC, TFIIF and TFIIE occupy similar positions to C34 and C82, respectively (Extended Data Fig. 8b). The 'E-wing' in TFIIE probably stabilizes the transcription bubble, in a similar manner to the C82 cleft loop. Therefore, the overall architecture of the downstream assembly is conserved, further confirming the hypothesis that general transcription factors have been stably incorporated into Pol III during evolution[10,39].

Moreover, our structures suggest that the 'TFIIE-like' heterotrimer combines the functions of both TFIIF and TFIIE, as the heterotrimer contributes DNA-binding winged-helix domains equivalent to those of TFIIE and TFIIF. However, similarity between the Pol II and Pol III PICs is restricted to a topological level, as the winged-helix domains in both systems do not superimpose. In particular, the C34 WH2 binds further downstream compared to the Tfg2 WH (Extended Data Fig. 8b), and is therefore likely to contribute directly to DNA melting by stabilizing the bubble.

We also note that the position of the Bdp1 SANT domain in the PIC is similar to that of TFIIA in the Pol II PIC, and both Bdp1 and TFIIF exchange β-strands with the protrusion (Extended Data Fig. 9), suggesting that Bdp1 combines the functions of TFIIF and TFIIA.

Despite the similar overall architecture of the Pol II and Pol III PICs, the way in which their respective promoter assemblies are formed is different. Pol III requires only TFIIIB, which forms very stable complexes on promoters and remains bound after each initiation event[40], with both arms of TFIIIB providing a scaffold for the inbuilt transcription-factor-like subcomplexes of Pol III. This enables Pol III to achieve very high initiation frequencies. By contrast, the minimal PIC of Pol II is transient and requires the factors TFIIB, TFIIE, TFIIF and TFIIH, of which TFIIB and preformed Pol II–TFIIF must rebind for each transcription initiation event[41–44]. This presumably enables tighter control, but results in a lower initiation frequency.

The architecture of the Pol I PIC that comprises the heterotrimeric core factor and Rrn3 is different to that of the Pol II and Pol III PICs, as in the Pol I PIC the upstream DNA already deeply penetrates the cleft[45–47] (Extended Data Fig. 8). In addition, whereas the cyclin folds of Brf1 and TFIIB contact the polymerase wall, the cyclin folds of Rrn7 do not (Extended Data Fig. 8c). The tandem WH domain of A49 may provide a 'downstream assembly' (Extended Data Fig. 8b) in Pol I, but it is disordered in most cryo-EM reconstructions[45–47] and lacks elements equivalent to the C82 cleft loop or the TFIIE E-wing. Therefore, the promoter assembly in Pol I is structurally, and probably mechanistically, different from those of Pol II and Pol III.

Model of promoter opening

We combined the structures reported here to obtain a model of the DNA melting process (Fig. 5, Supplementary Video 1). We also included a modelled intermediate (CC, open clamp), as we noticed that our CC structure is markedly different from those of yeast and human Pol II. In the Pol II system, DNA runs along the length of the polymerase and interacts with the jaw domain at the downstream end. In our structure, DNA can be traced only until bp −2, and projects away from Pol III. The clamp is in a closed state, the C82 cleft loop blocks access to the cleft, and the C34 WH1 and WH2 are disordered, presumably because they cannot occupy the same position as they do in the OC because they would clash with DNA (Fig. 4). It is therefore likely that the transition from the CC to the OC proceeds via an intermediate in which the clamp is open, similar to human Pol II and bacterial RNA polymerase[48,49]. We therefore modelled a CC, open-clamp intermediate based on the structure of apo Pol III[21]. In this model, the clamp and the cleft loop move upwards and lie on top of the closed DNA. The C34 winged helices can adopt similar positions as they do in the OC because the steric clash with DNA is removed, but they are in closer proximity to the DNA compared to their positions in the OC. This would trap DNA in the clamp, and clamp closing would enforce DNA melting by a steric clash of the C82 cleft loop and the DNA duplex, loading the template strand into the active site (Fig. 5, Supplementary Video 1). This process could be driven by the stabilization of the C34 WH2 in the OC conformation by Bdp1, C37 and the protrusion, as described earlier (see Fig. 2).

Our model is similar to that put forward for Pol II[49,50], as both use the coupled movements of the clamp, an extended loop (E-wing/C82 cleft loop) and winged-helix domains (heterotrimer/TFIIE–TFIIF) for promoter opening and nascent bubble stabilization. This suggests that the basic mechanisms of promoter opening are conserved between Pol II and Pol III.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 September; accepted 8 December 2017.

1. Goodfellow, S. J. & White, R. J. Regulation of RNA polymerase III transcription during mammalian cell growth. *Cell Cycle* **6,** 2323–2326 (2007).
2. White, R. J. RNA polymerases I and III, growth control and cancer. *Nat. Rev. Mol. Cell Biol.* **6,** 69–78 (2005).
3. Gouge, J. *et al.* Redox signaling by the RNA polymerase III TFIIB-related factor Brf2. *Cell* **163,** 1375–1387 (2015).
4. Johnson, S. A. S., Dubeau, L. & Johnson, D. L. Enhanced RNA polymerase III-dependent transcription is required for oncogenic transformation. *J. Biol. Chem.* **283,** 19184–19191 (2008).
5. Wu, L. *et al.* Novel small-molecule inhibitors of RNA polymerase III. *Eukaryot. Cell* **2,** 256–264 (2003).
6. Felton-Edkins, Z. A. & White, R. J. Multiple mechanisms contribute to the activation of RNA polymerase III transcription in cells transformed by papovaviruses. *J. Biol. Chem.* **277,** 48182–48191 (2002).
7. Dieci, G., Percudani, R., Giuliodori, S., Bottarelli, L. & Ottonello, S. TFIIIC-independent *in vitro* transcription of yeast tRNA genes. *J. Mol. Biol.* **299,** 601–613 (2000).
8. Kassavetis, G. A., Braun, B. R., Nguyen, L. H. & Geiduschek, E. P. *S. cerevisiae* TFIIIB is the transcription initiation factor proper of RNA polymerase III, while TFIIIA and TFIIIC are assembly factors. *Cell* **60,** 235–245 (1990).
9. Dieci, G. & Sentenac, A. Facilitated recycling pathway for RNA polymerase III. *Cell* **84,** 245–252 (1996).
10. Vannini, A. & Cramer, P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Mol. Cell* **45,** 439–446 (2012).
11. Keener, J., Josaitis, C. A., Dodd, J. A. & Nomura, M. Reconstitution of yeast RNA polymerase I transcription *in vitro* from purified components. TATA-binding protein is not required for basal transcription. *J. Biol. Chem.* **273,** 33795–33802 (1998).
12. Kassavetis, G. A., Driscoll, R. & Geiduschek, E. P. Mapping the principal interaction site of the Brf1 and Bdp1 subunits of *Saccharomyces cerevisiae* TFIIIB. *J. Biol. Chem.* **281,** 14321–14329 (2006).
13. Khoo, B., Brophy, B. & Jackson, S. P. Conserved functional domains of the RNA polymerase III general transcription factor BRF. *Genes Dev.* **8,** 2879–2890 (1994).
14. Gouge, J. *et al.* Molecular mechanisms of Bdp1 in TFIIIB assembly and RNA polymerase III transcription initiation. *Nat. Commun.* **8,** 130 (2017).
15. Ishiguro, A., Kassavetis, G. A. & Geiduschek, E. P. Essential roles of Bdp1, a subunit of RNA polymerase III initiation factor TFIIIB, in transcription and tRNA processing. *Mol. Cell. Biol.* **22,** 3264–3275 (2002).
16. Shah, S. M. A., Kumar, A., Geiduschek, E. P. & Kassavetis, G. A. Alignment of the B″ subunit of RNA polymerase III transcription factor IIIB in its promoter complex. *J. Biol. Chem.* **274,** 28736–28744 (1999).
17. Kumar, A., Kassavetis, G. A., Geiduschek, E. P., Hambalko, M. & Brent, C. J. Functional dissection of the B″ component of RNA polymerase III transcription factor IIIB: a scaffolding protein with multiple roles in assembly and initiation of transcription. *Mol. Cell. Biol.* **17,** 1868–1880 (1997).

18. Arimbasseri, A. G. & Maraia, R. J. Mechanism of transcription termination by RNA polymerase III utilizes a non-template strand sequence-specific signal element. *Mol. Cell* **58,** 1124–1132 (2015).
19. Rijal, K. & Maraia, R. J. RNA polymerase III mutants in TFIIFα-like C37 that cause terminator readthrough with no decrease in transcription output. *Nucleic Acids Res.* **41,** 139–155 (2013).
20. Kassavetis, G. A., Prakash, P. & Shim, E. The C53/C37 subcomplex of RNA polymerase III lies near the active site and participates in promoter opening. *J. Biol. Chem.* **285,** 2695–2706 (2010).
21. Hoffmann, N. A. *et al.* Molecular structures of unbound and transcribing RNA polymerase III. *Nature* **528,** 231–236 (2015).
22. Brun, I., Sentenac, A. & Werner, M. Dual role of the C34 subunit of RNA polymerase III in transcription initiation. *EMBO J.* **16,** 5730–5741 (1997).
23. Kassavetis, G. A., Soragni, E., Driscoll, R. & Geiduschek, E. P. Reconfiguring the connectivity of a multiprotein complex: fusions of yeast TATA-binding protein with Brf1, and the function of transcription factor IIIB. *Proc. Natl Acad. Sci. USA* **102,** 15406–15411 (2005).
24. Jakobi, A. J., Wilmanns, M. & Sachse, C. Model-based local density sharpening of cryo-EM maps. *eLife* **6,** e27131 (2017).
25. Wu, C. C., Lin, Y. C. & Chen, H. T. The TFIIF-like Rpc37/53 dimer lies at the center of a protein network to connect TFIIIC, Bdp1, and the RNA polymerase III active center. *Mol. Cell. Biol.* **31,** 2715–2728 (2011).
26. Hu, H.-L., Wu, C.-C., Lee, J.-C. & Chen, H.-T. A region of Bdp1 necessary for transcription initiation that is located within the RNA polymerase III active site cleft. *Mol. Cell. Biol.* **35,** 2831–2840 (2015).
27. Khoo, S.-K., Wu, C.-C., Lin, Y.-C., Lee, J.-C. & Chen, H.-T. Mapping the protein interaction network for TFIIB-related factor Brf1 in the RNA polymerase III preinitiation complex. *Mol. Cell. Biol.* **34,** 551–559 (2014).
28. Juo, Z. S., Kassavetis, G. A., Wang, J., Geiduschek, E. P. & Sigler, P. B. Crystal structure of a transcription factor IIIB core interface ternary complex. *Nature* **422,** 534–539 (2003).
29. Liao, Y., Moir, R. D. & Willis, I. M. Interactions of Brf1 peptides with the tetratricopeptide repeat-containing subunit of TFIIIC inhibit and promote preinitiation complex assembly. *Mol. Cell. Biol.* **26,** 5946–5956 (2006).
30. Male, G. *et al.* Architecture of TFIIIC and its role in RNA polymerase III pre-initiation complex assembly. *Nat. Commun.* **6,** 7387 (2015).
31. Kassavetis, G. A., Letts, G. A. & Geiduschek, E. P. A minimal RNA polymerase III transcription system. *EMBO J.* **18,** 5042–5051 (1999).
32. Kassavetis, G. A., Bardeleben, C., Kumar, A., Ramirez, E. & Geiduschek, E. P. Domains of the Brf component of RNA polymerase III transcription factor IIIB (TFIIIB): functions in assembly of TFIIIB–DNA complexes and recruitment of RNA polymerase to the promoter. *Mol. Cell. Biol.* **17,** 5299–5306 (1997).
33. Roy, K., Gabunilas, J., Gillespie, A., Ngo, D. & Chanfreau, G. F. Common genomic elements promote transcriptional and DNA replication roadblocks. *Genome Res.* **26,** 1363–1375 (2016).
34. Cloutier, T. E., Librizzi, M. D., Mollah, A. K. M. M., Brenowitz, M. & Willis, I. M. Kinetic trapping of DNA by transcription factor IIIB. *Proc. Natl Acad. Sci. USA* **98,** 9581–9586 (2001).
35. Kassavetis, G. A., Letts, G. A. & Geiduschek, E. P. The RNA polymerase III transcription initiation factor TFIIIB participates in two steps of promoter opening. *EMBO J.* **20,** 2823–2834 (2001).
36. Kassavetis, G. A., Blanco, J. A., Johnson, T. E. & Geiduschek, E. P. Formation of open and elongating transcription complexes by RNA polymerase III. *J. Mol. Biol.* **226,** 47–58 (1992).
37. He, Y., Fang, J., Taatjes, D. J. & Nogales, E. Structural visualization of key steps in human transcription initiation. *Nature* **495,** 481–486 (2013).
38. Sainsbury, S., Niesser, J. & Cramer, P. Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. *Nature* **493,** 437–440 (2013).
39. Carter, R. & Drouin, G. The increase in the number of subunits in eukaryotic RNA polymerase III relative to RNA polymerase II is due to the permanent recruitment of general transcription factors. *Mol. Biol. Evol.* **27,** 1035–1043 (2010).
40. Lassar, A. B., Martin, P. L. & Roeder, R. G. Transcription of class III genes: formation of preinitiation complexes. *Science* **222,** 740–748 (1983).
41. Yudkovsky, N., Ranish, J. A. & Hahn, S. A transcription reinitiation intermediate that is stabilized by activator. *Nature* **408,** 225–229 (2000).
42. Hahn, S. Activation and the role of reinitiation in the control of transcription by RNA polymerase II. *Cold Spring Harb. Symp. Quant. Biol.* **63,** 181–188 (1998).
43. Dieci, G. & Sentenac, A. Detours and shortcuts to transcription reinitiation. *Trends Biochem. Sci.* **28,** 202–209 (2003).
44. Rani, P. G., Ranish, J. A. & Hahn, S. RNA polymerase II (Pol II)–TFIIF and Pol II-mediator complexes: the major stable Pol II complexes and their activity in transcription initiation and reinitiation. *Mol. Cell. Biol.* **24,** 1709–1720 (2004).
45. Sadian, Y. *et al.* Structural insights into transcription initiation by yeast RNA polymerase I. *EMBO J.* **36,** 2698–2709 (2017).
46. Engel, C. *et al.* Structural basis of RNA polymerase I transcription initiation. *Cell* **169,** 120–131.e22 (2017).
47. Han, Y. *et al.* Structural mechanism of ATP-independent transcription initiation by RNA polymerase I. *eLife* **6,** e27414 (2017).
48. Feklistov, A. *et al.* RNA polymerase motions during promoter melting. *Science* **356,** 863–866 (2017).
49. He, Y. *et al.* Near-atomic resolution visualization of human transcription promoter opening. *Nature* **533,** 359–365 (2016).
50. Plaschka, C. *et al.* Transcription initiation complex structures elucidate DNA opening. *Nature* **533,** 353–358 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements C.W.M. and R.W. acknowledge support from the European Research Council Advanced Grant (ERC-2013-AdG340964-POL1PIC). M.K.V. acknowledges support from the EMBL International PhD program. H.K. acknowledges support from the European Union's Horizon 2020 program under the Marie Sklodowska-Curie grant agreement no. 703432. We thank F. Baudin for help with the transcription assay, A. J. Jakobi for extensive help with setting up LocScaling of cryo-EM maps and model refinement, T. Hoffmann and J. Pecar for setting up the high performance computing for RELION 2.0, and N. A. Hoffmann, J. Kosinski, L. Tafur and Y. Sadian for discussions. We thank A. Vannini for giving us access to the coordinates of PDB entry 5N9G before publication.

Author Contributions C.W.M. initiated and supervised the project. M.K.V. designed and carried out experiments, data processing and model building. H.K. helped in the initial stages of electron microscopy preparation. W.J.H.H. collected cryo-EM datasets. R.W. was responsible for yeast fermentation and helped with cloning. M.K.V. and C.W.M. prepared the manuscript with input from the other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to C.W.M. (christoph.mueller@embl.de).

Reviewer Information *Nature* thanks R. Maraia and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Protein expression and purification. Endogenous Pol III was purified as described[51] but exchanged into a buffer containing Li_2SO_4 instead of $(NH_4)_2SO_4$.

The Brf1–TBP plasmid[30] was transformed into BL21 Star (DE3) pRARE *Escherichia coli* cells. Expression cultures were grown at 37 °C in TB medium to an optical density at 600 nm of approximately 1.0, cooled (for 1 h at 4 °C) and induced with 50 μM IPTG overnight at 16 °C. Cells were pelleted for 5 min at 12,000*g* and re-suspended in 3 ml lysis buffer (1 M NaCl, 50 mM Tris pH 7.5, 2 mM β-mercaptoethanol, 20% glycerol, 10 μg ml^{-1} DNase I, 1 × protease inhibitors (SIGMAFAST protease inhibitor cocktail EDTA free), 30 mM imidazole, 2 mM $MgCl_2$) per gram of pellet. Cells were lysed in an Emulsiflex-C3 homogenizer and the lysate cleared by centrifugation for 1 h at 30,000*g*. The supernatant was incubated with 5 ml Ni-NTA resin (Qiagen) for 2 h. Beads were recovered and washed with 100 ml His-A buffer (1 M NaCl, 50 mM Tris pH 7.5, 2 mM β-mercaptoethanol, 5% glycerol, 30 mM imidazole) and 50 ml His-A low salt (similar to His-A but with 150 mM NaCl) and eluted with 50 ml His-B (200 mM NaCl, 2 mM β-mercaptoethanol, 5% glycerol, 300 mM imidazole). The eluate was loaded on a 5 ml HiTrap heparin column (GE Healthcare) pre-equilibrated in HepA buffer (similar to HisB but without imidazole). The column was washed with 6 column volumes containing 30% HepB (similar to HepA but with 1 M NaCl) and eluted with a linear gradient from 30% HepB to 70% HepB over 20 column volumes. Brf1–TBP eluted at around 600 mM NaCl. Peak fractions were concentrated and applied to a HiLoad 16/600 Superdex 200 size exclusion column equilibrated in 300 mM NaCl, 25 mM HEPES pH 7.5, 5 mM dithiothreitol and 5% glycerol. Purified Brf1–TBP was concentrated to approximately 6 mg ml^{-1}, then flash-frozen in liquid nitrogen and stored at −80 °C.

The Bdp1 plasmid[30] was transformed in BL21 Star (DE3) pRARE *E. coli* cells and grown in TB medium to an optical density at 600 nm of approximately 1.0, cooled and induced with 100 μM isopropyl β-D-1-thiogalactopyranoside overnight at 18 °C. Cell collection, lysis and Ni-NTA chromatography were performed as for Brf1–TBP. Eluted proteins were loaded onto a 5 ml heparin column pre-equilibrated in HepA. The column was washed with 6 column volumes of 20% HepB and eluted with a gradient from 20% HepB to 70% HepB over 30 column volumes. Peak fractions (approximately 520 mM NaCl) were digested with TEV protease overnight at 4 °C and incubated with 3 ml Ni-NTA for 1 h. The column was washed with 10 ml HepA containing 100 mM imidazole to recover cleaved Bdp1, which bound to Ni-NTA non-specifically. Bdp1 was finally purified by size exclusion chromatography as for Brf1–TBP (but in a buffer containing 150 mM NaCl) and concentrated to approximately 7 mg ml^{-1}. Aliquots were flash-frozen in liquid nitrogen and stored at −80 °C.

DNA oligonucleotides. For the preparation of preinitiation complexes, 81-nt long DNA scaffolds based on the U6 gene were used. The sequence contains the U6 promoter from −49 to +31 and, for the DNA–RNA hybrid, a 15-nt mismatch from −6 to +8 (template strand: 5′-CCAAATGTCCACGAAGGGTTACTTCGGCAACACATAGTTGCGAAAAAAACATTTATTTATAGTAGCCGAAAATAGTGGACG-3, non-template strand 5′-CGTCCACTATTTTCGGCTACTATAAATAAATGTTTTTTTCGCAGTCTATGCGGTTAACAGTAACCCTTCGTGGACATTTGG-3′, RNA 5′-GUUGCC-3′). For the DNA–RNA scaffold, we scrambled two positions in the template strand that base-pair with the RNA (wild-type sequence in template strand relative: +1-CAAGCG-+6, scrambled sequence: +1-CAACGG-+6) in order to remove an alternative complementary site to which the RNA could bind. The closed DNA scaffold contains the wild-type sequence (template strand: 5′-CCAAATGTCCACGAAGGGTTACTTCGCGAACACATAGTTGCGAAAAAAACATTTATTTATAGTAGCCGAAAATAGTGGACG-3′).

Duplex DNA was generated by mixing single-stranded DNA oligonucleotides in H_2O and heating to 95 °C for 10 min. The reaction was cooled to 20 °C at a rate of 1.5 °C per min. For the DNA–RNA hybrid, RNA was added to the DNA duplex and heated to 40 °C for 10 min and then cooled to 4 °C.

Assembly of preinitiation complexes. For *in vitro* reconstitution of the Pol III–PIC we resorted to a fusion construct of Brf1 and TBP (Brf1–TBP) that has been shown to substitute Brf1 and TBP function *in vitro* and *in vivo*[23]. For the ITC, we preincubated Pol III with the ITC scaffold, which ensures the positioning of Pol III with the correct polarity on the transcription bubble due to binding of the DNA–RNA hybrid in the active site. 300 μg of Pol III was incubated with a 1.1× excess of DNA–RNA scaffold for 20 min and Brf1–TBP and Bdp1 were added at a 3× excess for 1 h on ice. Samples were diluted to 0.5 mg ml^{-1} (calculated for Pol III) in cross-linking buffer (100 mM Li_2SO_4, 15 mM HEPES pH 7.5, 10 mM dithiothreitol, 5 mM $MgCl_2$, 0.05% glutaraldehyde) and cross-linked on ice for 30 min. Cross-linking was quenched by addition of 40 mM Tris pH 7.5.

For the closed complex, the order was reversed and Brf1–TBP and Bdp1 were incubated with DNA for 20 min to position TFIIIB on the TATA box before Pol III was added. Cross-linked samples were concentrated in spin concentrators (Amicon Ultra, 500 μl, 30 kDa cutoff) and applied to a Superose 6 increase 3.2/300 column equilibrated in EM buffer (150 mM Li_2SO_4, 15 mM HEPES pH 7.5, 10 mM dithiothreitol, 5 mM $MgCl_2$). Fifty-microlitre fractions were collected and the peak fraction was used for cryo-EM grid preparation.

Electron microscopy. Cryo-grids were prepared with a Vitrobot IV set to 100% humidity and 4 °C. Quantifoil 200 mesh Cu 2/1 grids were glow discharged in a Pelco EasyGlow glow discharger and 2.5 μl of sample was applied (blotting parameters: wait time 15 s, blot force 4, blot time 4 s) and plunge-frozen in liquid ethane.

Micrographs were acquired on a Titan Krios operated at 300 keV equipped with a Gatan Quantum energy filter and a K2 Summit direct detector. The detector was operated in super resolution mode at 105,000× magnification and a calibrated physical pixel size of 1.35 Å. Data collection parameters and dataset sizes are shown in Extended Data Table 1.

For all datasets, frame alignment and dose weighting were performed with MotionCor2[52] and contrast transfer function parameters estimated with Gctf[53]. Particle picking and classification were performed with RELION 2.0[54].

For the ITC dataset, 773,000 auto-picked particles were extracted and binned 4 times to reduce computational costs (Extended Data Fig. 2). Particles were subjected to 3D classification using a 60-Å low-pass filtered model of apo Pol III (PDB: 5FJ9) without prior 2D classification. The major of the 4 classes contained 60% or 464,000 particles and showed clear density corresponding to TFIIIB and promoter DNA. Particles of that class were unbinned, extracted in 300-pixel boxes and refined. The resulting volumes showed clearly defined secondary structure elements for TFIIIB, but at a lower threshold compared to the Pol III core. Hence, we performed masked classification using a mask that covers TFIIIB ('TFIIIB mask #1'), upstream DNA and C34 WH1 and WH2. Out of the 3 classes, a smaller class with 79,000 particles (10.2% of the autopicked particles) showed very well defined TFIIIB density at the same threshold as the Pol III core. The class had clear density for downstream DNA, but at a lower threshold than upstream DNA. Therefore we performed classification using a mask on downstream DNA. This yielded the $OC_{\Delta downstream\text{-}1}$ class (38,000 particles, 4.9%) and the ITC class (29,000 particles, 3.8%) maps.

We collected two datasets for the closed complex. Both were initially processed separately as the ITC dataset until the first masked classification using TFIIIB mask no. 1 (Extended Data Fig. 2). We noticed that a minor class in both datasets had reduced density for C34 WH1 and WH2 and a different path of the upstream DNA. We pooled all particles that had strong TFIIIB density in both CC datasets (particle set 'CC_joined'), which contained 226,000 particles and was refined to 3.4 Å ('Pol III PIC_joined' map). We classified the CC_joined particle set with a global mask (250 Å diameter) and noticed that one class showed stronger DNA density in the cleft as well as additional density adjacent to Pol III. Reclassification of these particles with a 400-Å mask revealed a subset of dimers (3% of CC_joined); in the dimers, two Pol III molecules bind to the same DNA molecule with their clefts (Extended Data Fig. 2). We excluded dimers and pooled all other classes, which we classified using a mask that covered C34 WH1 and WH2. From this we obtained two classes (142,000 particles, 67% of CC_joined) that exhibited upstream DNA and C34 density as in the ITC/OC1 maps that were pooled was refined to 3.5 Å (not shown), and a minor CC class that had very weak C34 WH1 and WH2 density and showed the path of the upstream DNA (68,000 particles, 30% of CC_joined). The CC class was further classified using a mask covering TFIIIB and upstream DNA as seen in the CC as well as C34 WH1 and WH2, which yielded the CC2 (34,000 particles, 15% of CC_joined) and the CC1 reconstructions (19,000 particles, 8.2% of CC_joined).

We further classified the 3.5 Å OC class using a mask on downstream DNA, which yielded the OC (62,000 particles, 29% of CC_joined) and a map with disordered downstream DNA $OC_{\Delta downstream\text{-}2}$ (79,000 particles, 35.9% of CC_joined).

Model building. We constructed an initial model by combining the structures of elongating Pol III (PDB: 5FJ8) and the Brf1–TBP core crystal structure (PDB: 1NGM), encompassing TBP residues 61–240 and Brf1 residues 437–506. Next, we used Phyre2[55] to calculate homology models of Brf1 (residues 70–270), Brf1 (residues 1–40) and the Bdp1 SANT domain (residues 416–464) and fitted these as well. C34 WH1 and WH2 homology models were based on the NMR structure of the corresponding mouse domains (PDB: 2DK8, 2DK5) and generated with Modeller[56] and placed into the density. Then we extended the Bdp1 structure by manual modelling in Coot[57]. Towards the C terminus, we extended the SANT domain with two additional helices that were predicted on a secondary-structure level and noticed that a prominent density running in parallel with the Brf1 homology domain II fitted the predicted coiled-coil in Bdp1 perfectly. Towards the N terminus, we manually built the linker and tether regions of Bdp1. Building was aided by secondary structure predictions and bulky side chains that were visible in the cryo-EM density (Extended Data Fig. 4). During the preparation of this manuscript, the crystal structure of human TBP–Brf2 (residues 64–407)–Bdp1 (residues 286–407) became available[14], giving us additional confidence in assigning the N-terminal density of the Bdp1 SANT domain. Finally, the prominent density running underneath the DNA was assigned to Bdp1 residues 275–320, based on available protein–DNA cross-links[16,58] that placed a region of Bdp1 on the opposite side of TBP and was mapped to Bdp1 residues 299–315. Reported photo-crosslinks of Bdp1 K281 to Brf1 and Bdp1 residues 291–295 to the C128 protrusion gave additional confidence[26]. Finally, our cryo-EM density in the 3.7 Å map showed side chain density in this region that enabled us to obtain the sequence register (Extended Data Fig. 5). We also used the 3.4 Å 'Pol III PIC_joined' map for model building as it showed improved side-chain densities that were especially helpful for building the Bdp1 ER I and Bdp1 tether region.

Local amplitude scaling (LocScale). LocScale maps were calculated as described[24]. In brief, unsharpened and unfiltered cryo-EM maps were scaled against simulated model maps using a rolling window corresponding to seven times the average resolution of the cryo-EM map (19 voxels (25.7 Å) for the OC map, 22 voxels (29.7 Å) for the ITC and CC2 map, and 25 voxels (33.8 Å) for the CC1 map). Model maps were simulated from full-sidechain models refined against cryo-EM maps that were sharpened using the relion_postprocess program. The $d_{\min}$ parameter of the LocScale program was set to the Nyquist frequency (2.7 Å).

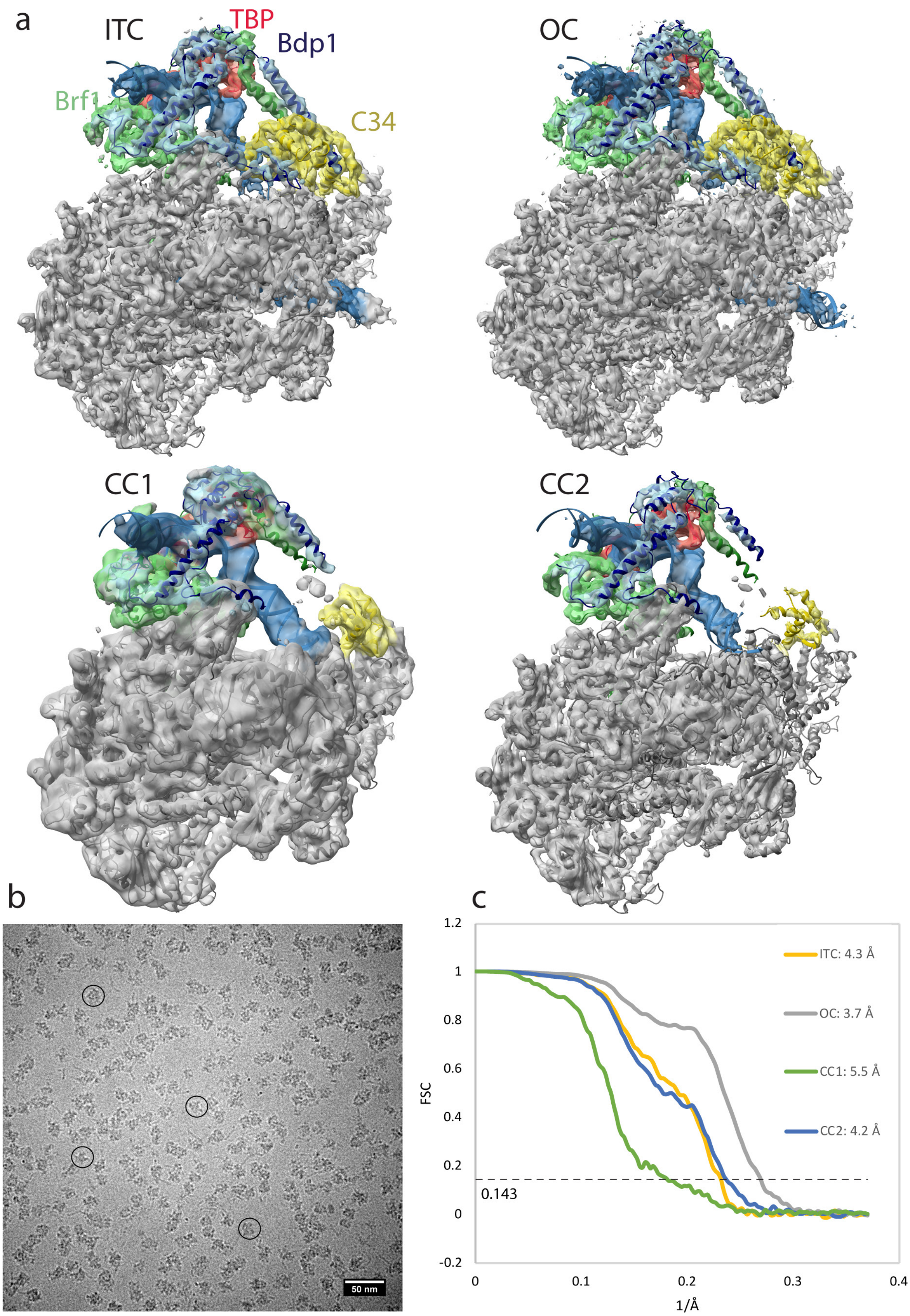
Model refinement and validation. Models were refined against respective cryo-EM maps which were B-factor sharpened with relion_postprocess. We used a refinement strategy essentially as described previously[21,59] based on PHENIX[60] libraries. Geometry statistics were calculated with MolProbity[61]. The local resolution was calculated with blocres[62] using a box size of 20 voxels and a Fourier shell correlation cutoff of 0.5. Figures were prepared with UCSF Chimera[63] and PyMOL[64].

Modelling of the open-clamp, closed complex structure. The open-clamp, CC model was obtained by combining the structure of open-clamp apo Pol III (PDB: 5FJA) with the DNA of the human Pol II PIC in its closed state (PDB: 5IYA); this DNA has a much closer resemblance to that of the Pol III OC than does DNA of the yeast Pol II OC. The position of C34 was obtained by superimposing C34 from the Pol III OC model (this work) onto the open-clamp structure and aligning the WH3 domains. In this model there are only minor clashes between DNA and the Rpb5 subunit in the jaw, which could easily be accommodated by a slightly different curvature of the DNA.

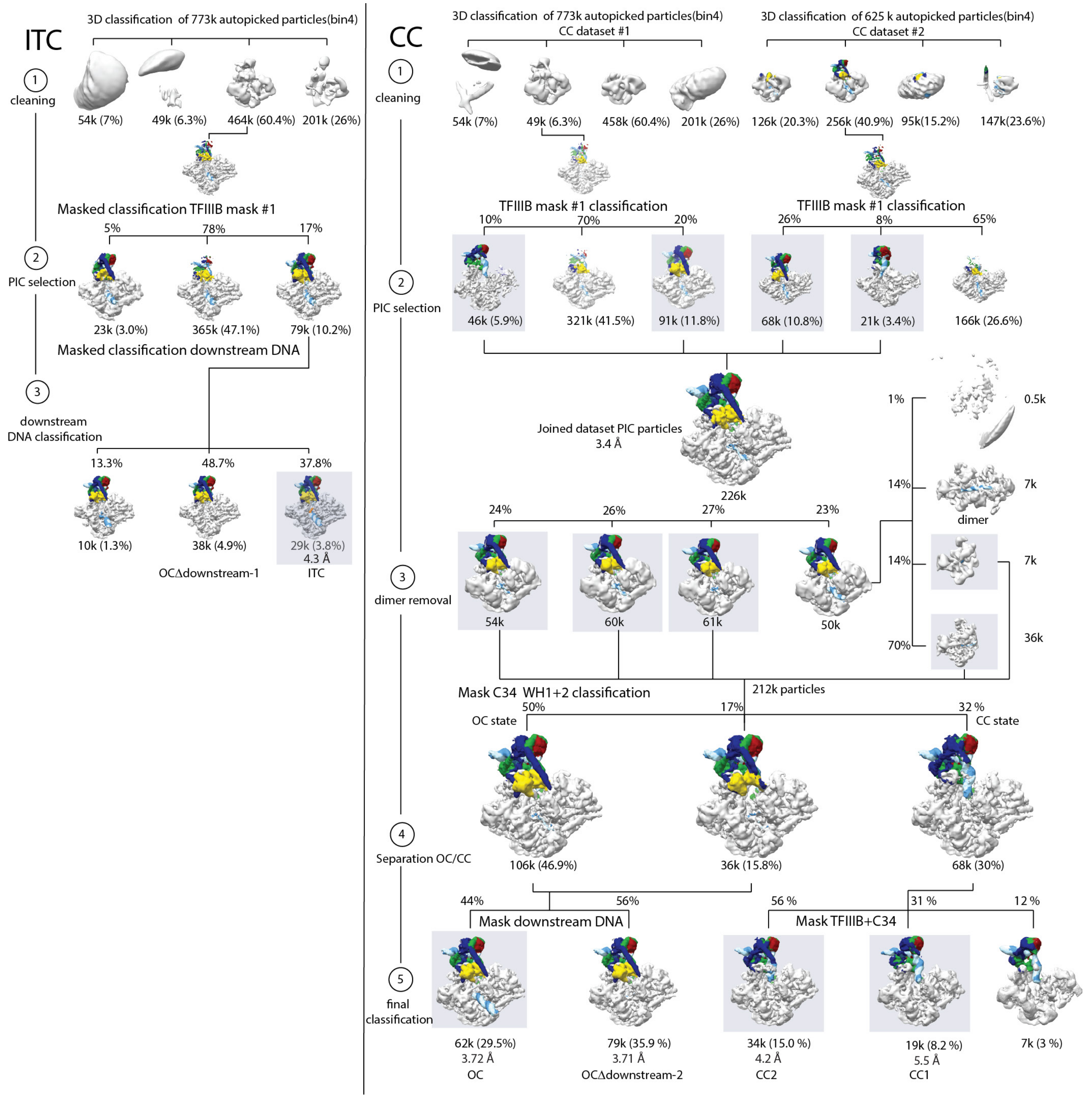
RNA extension assay. The RNA extension assay was performed with the DNA oligonucleotides and 6-mer RNA described (ITC scaffold). RNA was first radioactively labelled with ^{32}P using T4 polynucleotide kinase and purified over a 15% denaturing urea–polyacrylamide gel. The ITC scaffold was annealed as described above but with the ^{32}P-labelled RNA, and 2 pmol scaffold was preincubated with 4 pmol of Pol III for 20 min at 20 °C, followed by incubation with 12 pmol of TFIIIB for 20 min at 20 °C in 20 mM Tris pH 7.5, 200 mM NaCl, 10 mM dithiothreitol and 10 mM $MgCl_2$. RNA elongation was initiated by the addition of 1 mM ATP, GTP and UTP. After incubation for 10 min at 28 °C, RNA extension was stopped by the addition of 0.1% SDS, 30 mM EDTA. After phenol extraction and ethanol precipitation, the resulting ^{32}P-labelled RNA products were separated on a denaturing 15% polyacrylamide–urea gel. RNA bands were detected on an imaging plate (Fujifilm) using a Typhon FLA9500 phosphoimager. The digital image was cropped and contrast was adjusted with the 'levels' tool in Photoshop CS6 v13.0.1.

Data availability. Cryo-EM maps of the Pol III ITC, Pol III OC, Pol III CC1 and Pol III CC2, and an intermediate 3.4 Å OC map that aided model building (Pol III PIC_joined, see Methods), have been deposited in the Electron Microscopy Data Bank (EMDB) under accession codes EMD-4181 (Pol III ITC), EMD-4180 (Pol III OC), EMD-4182 (Pol III CC1) and EMD-4183 (Pol III CC2) and EMD-4184 (Pol III PIC_joined). The coordinates of the corresponding atomic models have been deposited in the Protein Data Bank (PDB) under accession codes 6F41 (Pol III ITC), 6F40 (Pol III OC1), 6F42 (Pol III CC1) and 6F44 (Pol III CC2)

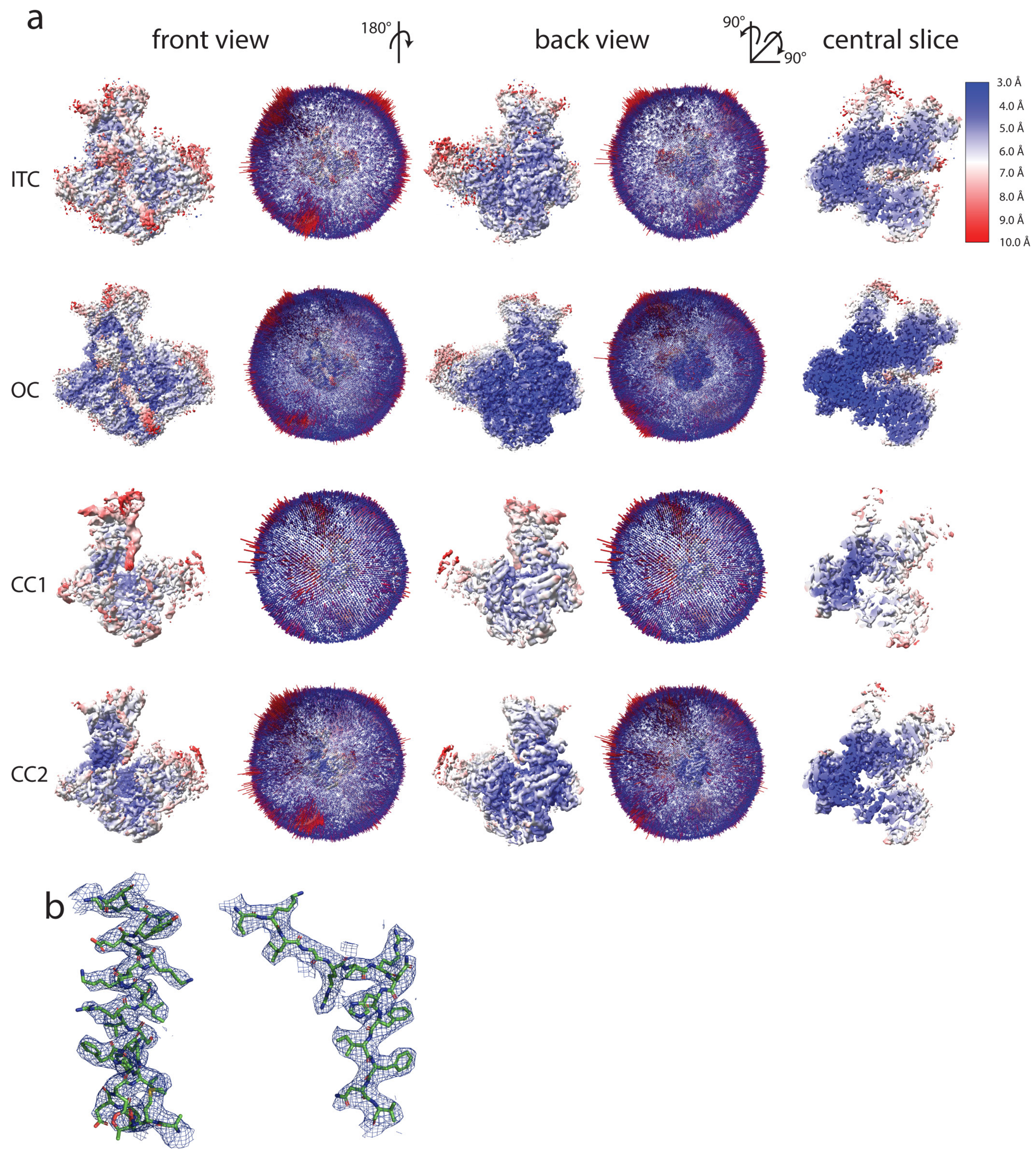
51. Moreno-Morcillo, M. *et al.* Solving the RNA polymerase I structural puzzle. *Acta Crystallogr. D* **70,** 2570–2582 (2014).
52. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14,** 331–332 (2017).
53. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193,** 1–12 (2016).
54. Kimanius, D., Forsberg, B. O., Scheres, S. H. W. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5,** e18722 (2016).
55. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10,** 845–858 (2015).
56. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815 (1993).
57. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
58. Kang, J. J., Kang, Y. S. & Stumph, W. E. TFIIIB subunit locations on U6 gene promoter DNA mapped by site-specific protein–DNA photo-cross-linking. *FEBS Lett.* **590,** 1488–1497 (2016).
59. Tafur, L. *et al.* Molecular structures of transcribing RNA polymerase I. *Mol. Cell* **64,** 1135–1143 (2016).
60. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
61. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
62. Heymann, J. B. & Belnap, D. M. Bsoft: Image processing and molecular modeling for electron microscopy. *J. Struct. Biol.* **157,** 3–18 (2007).
63. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
64. The PyMOL Molecular Graphics System v.1.8 (Schrödinger, 2015).



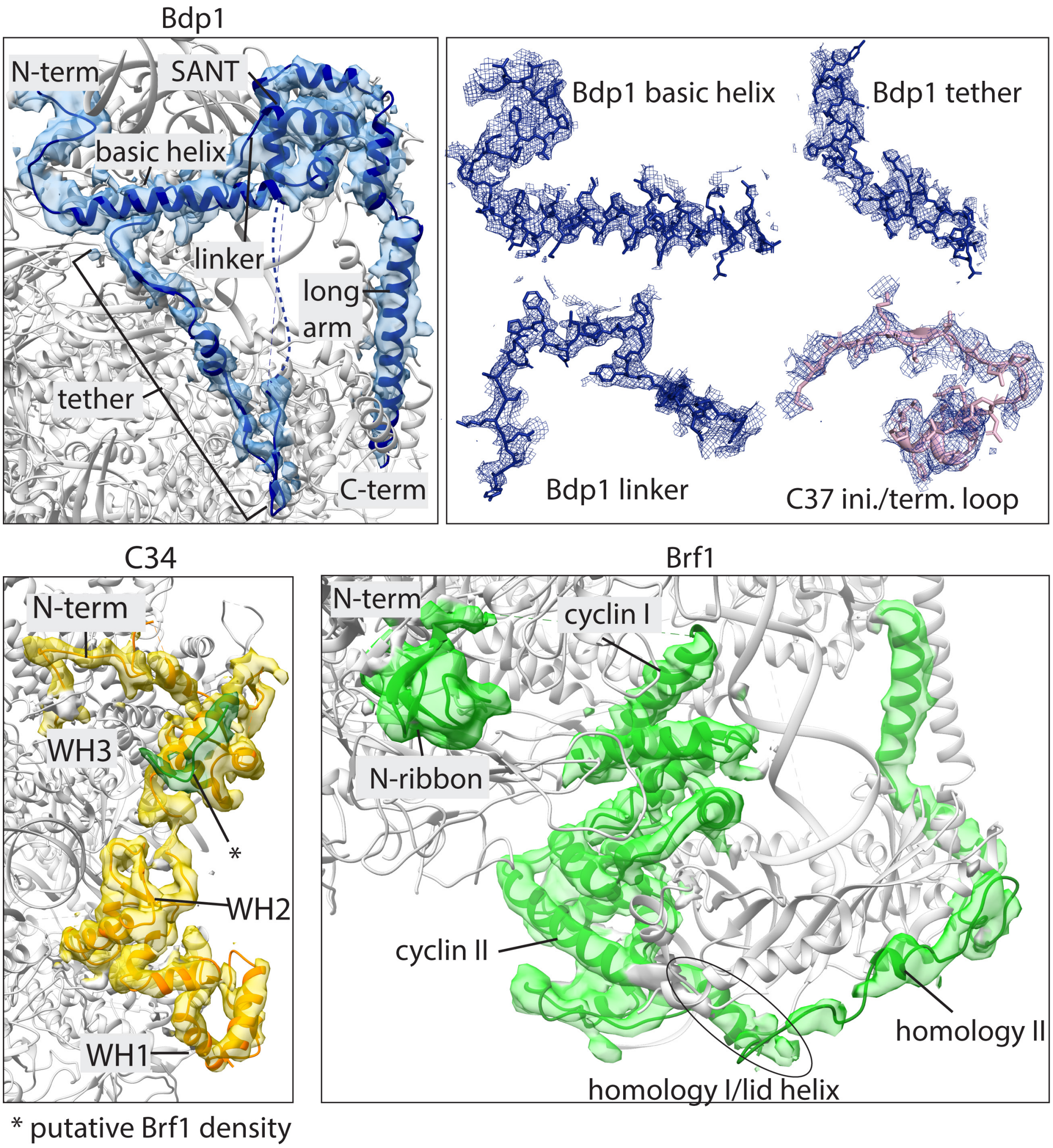
Extended Data Figure 1 | Cryo-EM analysis of the Pol III PIC. **a**, Cryo-EM densities sharpened with relion_postprocess of the Pol III ITC, OC, CC1 and CC2 maps reported here superimposed with the corresponding models. **b**, Typical micrograph of the Pol III PIC. **c**, Fourier-shell correlation curves for the different cryo-EM maps, as reported by the relion_postprocess program.



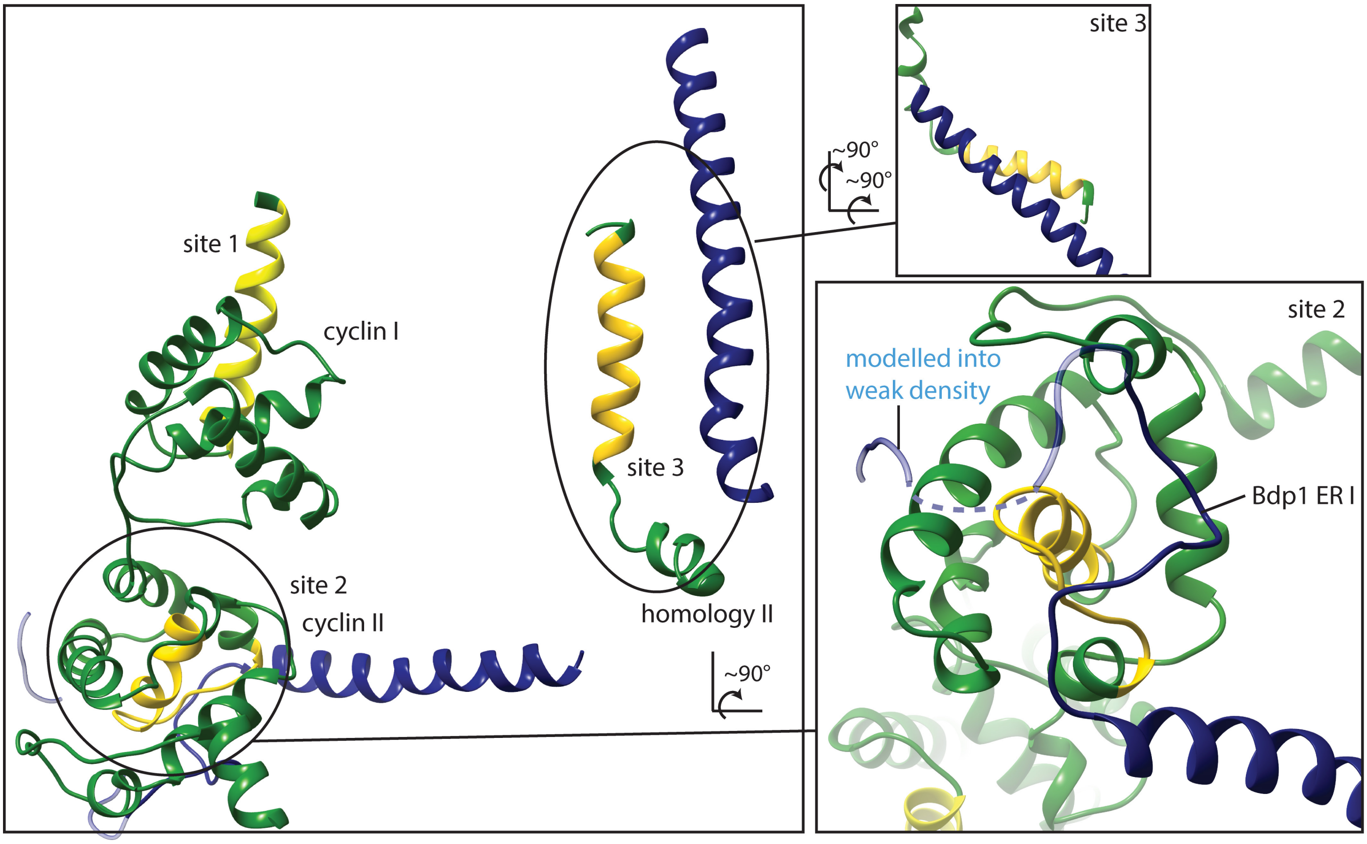
Extended Data Figure 2 | Classification strategies for cryo-EM datasets. Left, ITC dataset. The ITC dataset was cleaned (step 1) by classification in 3D (binned 4 times) and PIC particles were selected using a mask on TFIIIB (step 2). Finally, classification using a mask on downstream DNA yielded the ITC and the OC$_{\Delta\text{downstream}-1}$ maps (step 3). Right, CC datasets. Two CC datasets were individually cleaned (step 1) by classification in 3D (binned 4 times) and PIC particles were selected using a mask on TFIIIB (step 2). PIC particles from both datasets were combined and classified without masking to remove a small subset of dimers (step 3). Classification using a mask on C34 WH1 and WH2 yielded the OC and CC populations (step 4) which were subsequently classified into CC1, CC2, OC and OC$_{\Delta\text{downstream}-2}$ (step 5). The majority of particles from the CC datasets are in the OC state (67% of PIC particles, versus 32% in the CC states, Extended Data Fig. 2), showing that our TFIIIB–Pol III complex is active in promoter opening. Focused classification on downstream DNA in the ITC and CC datasets gave rise to reconstructions which show OC-like upstream DNA but lack downstream DNA (OC$_{\Delta\text{downstream}}$). This may suggest the existence of an initiation intermediate in which downstream DNA is mobile and becomes ordered only in a later stage of the transcription cycle, as has also been described for bacterial RNA polymerase[48]. An alternative explanation for the lack of density in the OC$_{\Delta\text{downstream}}$ map derived from the CC scaffold lies in the pseudo-symmetric nature of the U6 TATA box[28]. Because we used a wild-type promoter sequence, a subset of particles may have bound the DNA scaffold with reverse polarity, making the ‘downstream’ DNA too short to be stabilized in the cleft. However, we favour the former explanation, as we also observe OC$_{\Delta\text{downstream}}$ in the ITC dataset, in which the polarity is defined by incubating Pol III with the preformed transcription bubble.



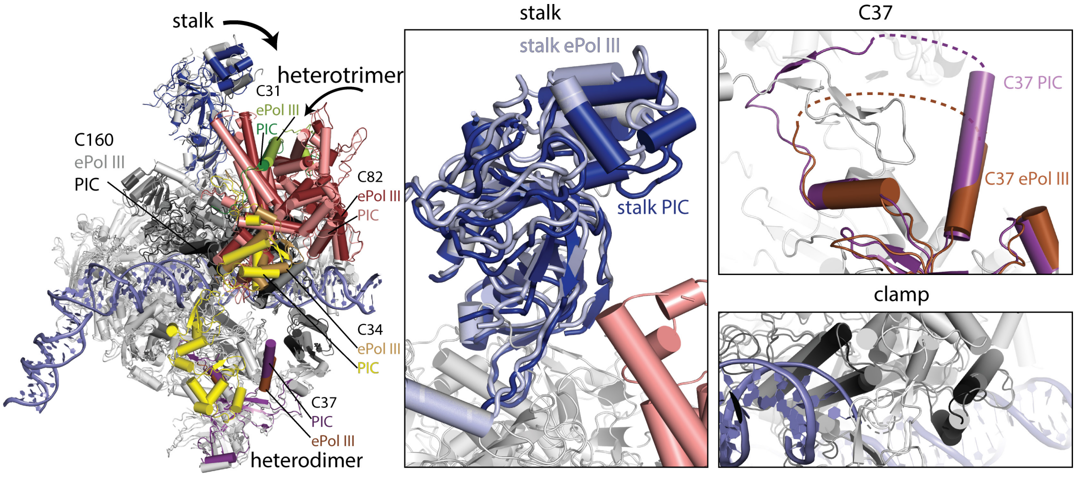
Extended Data Figure 3 | Local resolution, Euler angular distribution and side chain densities. a, Cryo-EM maps coloured by local resolution from 3 Å (blue) to 10 Å (red). All maps are shown on the same contour level. The central slices show that the local resolution degrades more strongly in the CC2 map compared to the ITC map, although they have comparable overall resolutions. Euler angular distribution plotting shows a good angular coverage without a dominating preferred orientation in all reconstructions. **b**, Examples of helical and β-strand densities in Pol III subunit AC40 in the 3.7 Å OC map.



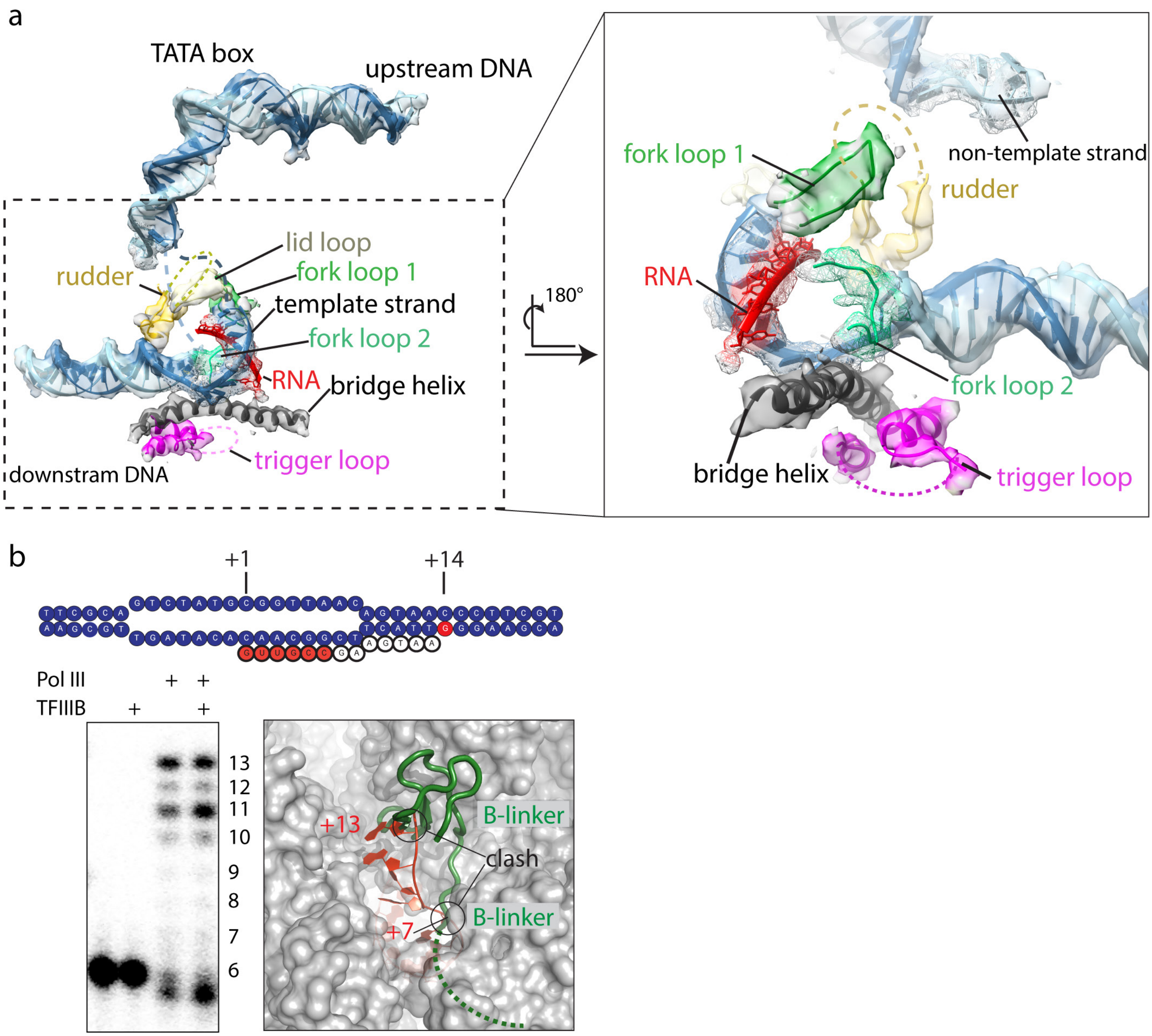
Extended Data Figure 4 | Cryo-EM density for newly modelled regions. Cryo-EM densities in the ITC map of Bdp1 (top), C34 WH1 and WH2 (bottom left), and Brf1 (bottom right) after amplitude scaling contoured at a level that allows visualization of most residues. Top right, side-chain models and cryo-EM densities for *de novo* built regions.



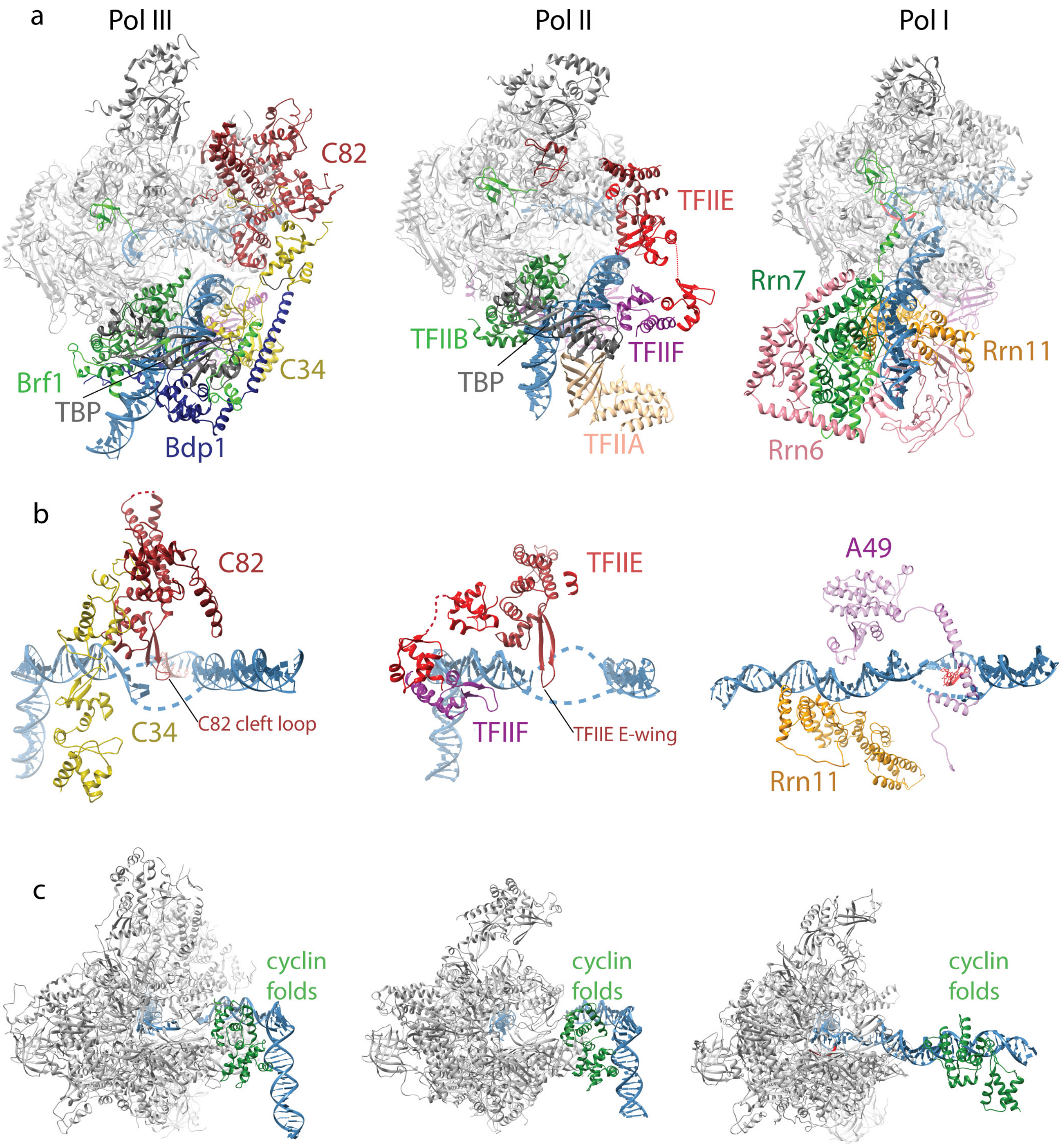
Extended Data Figure 5 | Bdp1 partially masks binding sites for TFIIIC in Brf1. Brf1 regions that have been mapped to interact with TFIIIC are shown in yellow[29]. Sites 2 and 3 are partially buried by Bdp1 in the PIC. Right, close-up views of site 3 (top) and site 2 (bottom). The parts of Bdp1 shown as transparent are those that could be built but were not included in the final model owing to lack of sequence register and weak density. Bdp1 appears to compete for the same binding sites on Brf1 as TFIIIC, suggesting that the correct assembly of TFIIIB might trigger the conformational rearrangements in the assembly factor TFIIIC that render the complex initiation-competent.



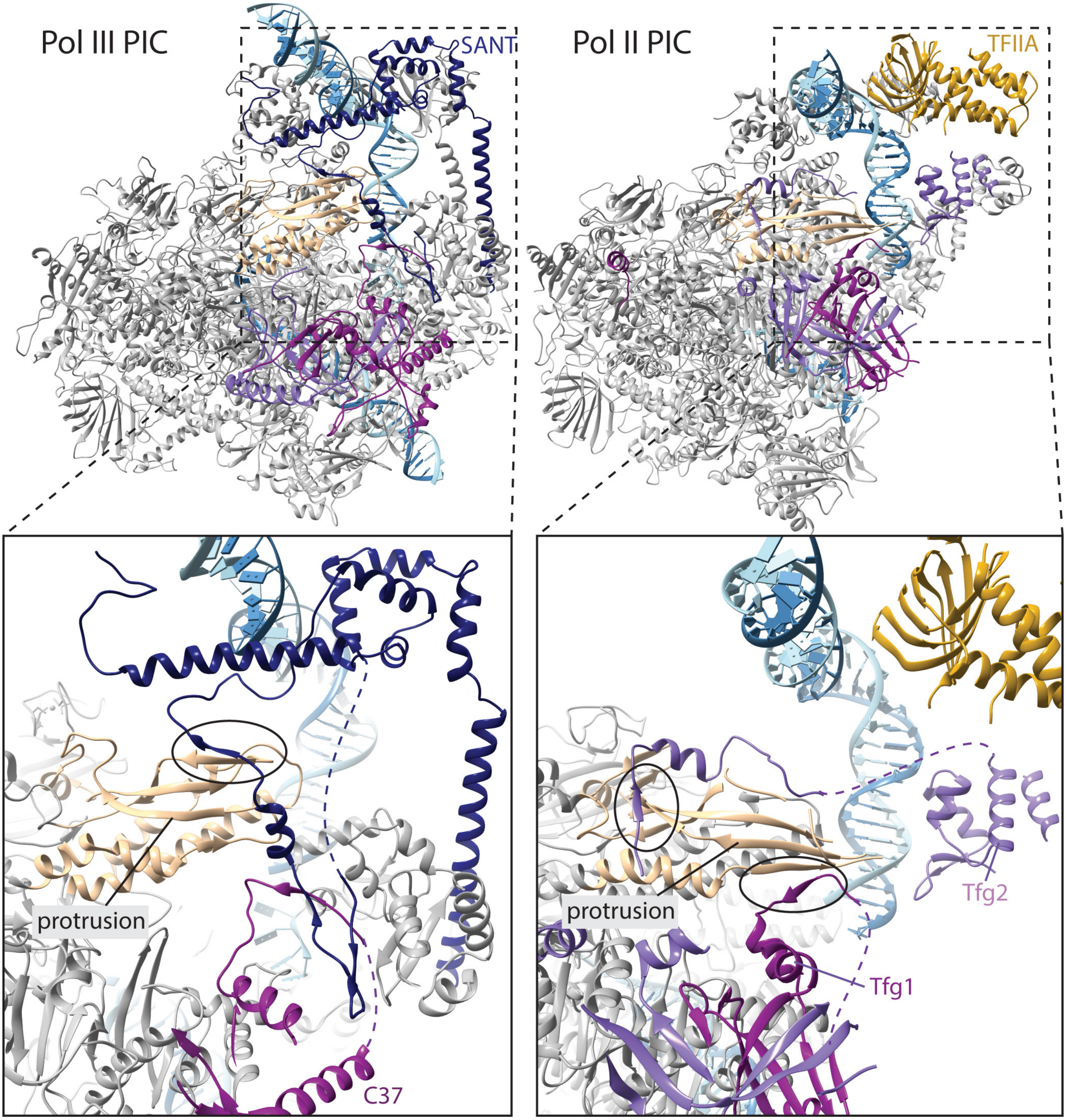
Extended Data Figure 6 | Comparison between Pol III in the OC and elongating states. Superposition of elongating Pol III (ePol, PDB: 5FJ8) and Pol III in the OC. In the OC, the heterotrimer moves towards upstream DNA and the C34 WH1 and WH2 domains become ordered. The stalk moves towards the heterotrimer, and the clamp moves to slightly close the cleft. The C37 initiation/termination loop becomes partially ordered to interact with the Bdp1 tether and C34 WH1 and WH2.



Extended Data Figure 7 | Active site and nucleic acid density in the ITC. a, Cryo-EM density in the ITC map after amplitude scaling of active-site elements and nucleic acids. Elements contoured at a lower threshold are shown as mesh. Active-site elements are labelled. The rudder and trigger loops are disordered. The poor density of RNA suggests flexibility or partial occupancy due to dissociation during sample preparation or cleavage by C11. **b**, RNA extension assay of the ^{32}P-labelled 6-nt RNA in the absence of CTP, showing that our preparation is active in transcription using our ITC scaffold. Addition of TFIIIB leads to a stronger accumulation of 11-nt RNA. Modelling of an elongated RNA oligonucleotide based on RNA polymerase I elongation complex 1 (PDB: 5M5X) shows that RNA clashes at position +7 with the beginning of the flexible B-linker, and at position +13 with the B-ribbon. Accumulation of 11-nt RNA in the presence of TFIIIB suggests that the RNA in the Pol III PIC takes a different path compared to that in the Pol I elongation complex, and clashes with the B-ribbon at position +12, requiring the release of TFIIIB to enter elongation. The experiment was repeated three times independently with the same results.



Extended Data Figure 8 | Comparison of Pol III, Pol II and Pol I PICs. **a**, Ribbon diagrams of yeast Pol III, Pol II (PDB: 5FYW) and Pol I (PDB: 5W65) PICs. TFIIF and TFIIE occupy similar positions to the C53–C37 heterodimer and the C82–C34–C31 heterotrimer, respectively. The convex surface of TBP is closely contacted by Brf1 homology domain II in TFIIIB, but is accessible in the Pol II PIC, whereas TBP is entirely absent from available structures of Pol I PIC. This might explain the strict requirement for TBP in the Pol III system, in contrast to Pol II and Pol I. **b**, Close-up view of the downstream promoter assembly, showing C82 WH3–WH4, C34 WH1–WH3 (left), the WH domains of TFIIF and TFIIE (middle) and the A49 tandem WH9 and TPR of Rrn11 (right). Whereas C82–C34 and TFIIF–TFIIE form structurally similar downstream promoter assemblies using WH domains and the C82 'cleft loop' and TFIIE 'E-wing' to contact the upstream bubble edge, the Pol I core factor forms a structurally different assembly in which the A49 tandem WH does not contact the upstream bubble edge in the same way. **c**, Comparison of the cyclin folds in Brf1, TFIIB and Rrn7. The cyclin folds in Brf1 and TFIIB occupy similar positions and contact the polymerase wall, whereas the cyclin folds in Rrn7 do not.



Extended Data Figure 9 | Comparison between Bdp1 in the Pol III PIC and TFIIA and TFIIF in the Pol II PIC. The Bdp1 SANT domain is located at a similar position to TFIIA. Parts of Bdp1 resemble TFIIF subunit Tfg1, although no sequence similarity is detectable. Both interact with the Pol protrusion by adding a β-strand to it (although at different ends of the protrusion β-sheet) and folding into a short helix along the face of the protrusion. The path of the C37 initiation/termination loop is also similar to that of Tfg1. The second subunit of TFIIF, Tfg2, also adds a β-strand to the Pol II protrusion. Brf1 and TBP were omitted for clarity.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	#1 ITC (EMD-4181) (PDB 6F41)	#2 OC (EMD-4180) (PDB 6F40)	#4 CC1 (EMD-4182) (PDB 6F42)	#3 CC2 (EMD-4183) (PDB 6F44)
Data collection and processing				
Magnification	105,000	105,000	105,000	105,000
Voltage (kV)	300	300	300	300
Electron exposure (e^-/Å^2)[1]	61.3	60.0/61.8	60.0/61.8	60.0/61.8
Defocus range (μm)	-1 to -3	-0.5 to -4	-0.5 to -4	-0.5 to -4
Pixel size (Å)	1.35	1.35	1.35	1.35
Symmetry imposed	n/a	n/a	n/a	n/a
Initial particle images (no.)	472,519	714,312	714,312	714,312
Final particle images (no.)	29,951	62,751	18,760	34,176
Map resolution (Å)	4.3	3.7	5.5	4.2
FSC threshold	0.143	0.143	0.143	0.143
Map resolution range (Å)	3.9-9.0	3.3-7.5	4.0-9.0	3.7-9.0
Refinement				
Initial model used (PDB code)	5FJ8	5FJ8	5FJ8	5FJ8
Model resolution (Å)	4.3	3.7	5.5	4.2
Map sharpening *B* factor (Å^2)	142	130	100	90
Model composition				
Non-hydrogen atoms	48,614	47,912	45,594	45,573
Protein residues	5,721	5,721	5,515	5,512
Ligands	0	0	0	0
***B* factors (Å^2)**				
Protein	125.93	57.53	254.30	121.87
Ligand	n/a	n/a	n/a	n/a
R.m.s. deviations				
Bond lengths (Å)	0.01	0.01	0.01	0.01
Bond angles (°)	1.03	1.04	1.03	1.03
Validation				
MolProbity score	2.12	2.06	2.18	2.07
Clashscore	8.71	8.23	11.19	8.77
Poor rotamers (%)	1.12	0.96	1.07	0.81
Ramachandran plot				
Favored (%)	87.37	87.17	88.35	88.01
Allowed (%)	12.59	12.72	11.56	11.92
Disallowed (%)	0.04	0.11	0.09	0.07

[1] Structures of OC, CC1 and CC2 were determined from two datasets. The electron dose is given for both datasets.

doi:10.1038/nature25441

Structural basis of RNA polymerase III transcription initiation

Guillermo Abascal-Palacios[1], Ewan Phillip Ramsay[1], Fabienne Beuron[1], Edward Morris[1] & Alessandro Vannini[1]

RNA polymerase (Pol) III transcribes essential non-coding RNAs, including the entire pool of transfer RNAs, the 5S ribosomal RNA and the U6 spliceosomal RNA, and is often deregulated in cancer cells. The initiation of gene transcription by Pol III requires the activity of the transcription factor TFIIIB to form a transcriptionally active Pol III preinitiation complex (PIC). Here we present electron microscopy reconstructions of Pol III PICs at 3.4–4.0 Å and a reconstruction of unbound apo-Pol III at 3.1 Å. TFIIIB fully encircles the DNA and restructures Pol III. In particular, binding of the TFIIIB subunit Bdp1 rearranges the Pol III-specific subunits C37 and C34, thereby promoting DNA opening. The unwound DNA directly contacts both sides of the Pol III cleft. Topologically, the Pol III PIC resembles the Pol II PIC, whereas the Pol I PIC is more divergent. The structures presented unravel the molecular mechanisms underlying the first steps of Pol III transcription and also the general conserved mechanisms of gene transcription initiation.

In the eukaryotic nucleus, RNA Pol III catalyses the DNA-dependent synthesis of short RNAs that are essential for cellular functions, such as tRNAs, the 5S rRNA and the U6 spliceosomal small nuclear RNA (snRNA). Pol III transcriptional output is a key determinant of cellular and organismal growth[1] and misregulated Pol III activity can cause several diseases, including cancer[2–5].

Pol III is predominantly regulated at the level of transcription initiation. The assembly of a PIC, which is formed when Pol III is recruited to its target genes by a specific set of general transcription factors (GTFs), is a highly regulated process[6–8]. The TFIIIB complex is a multi-subunit GTF that is ubiquitously required at Pol III-transcribed genes *in vivo*[9–11] and is sufficient to form a transcriptionally competent Pol III PIC *in vitro*[12–14]. TFIIIB consists of the TATA-box binding protein (TBP), B-related factor 1 (Brf1), which is functionally related to the Pol II paralogue TFIIB[15], and the B double prime (Bdp1) subunit[16].

The 17-subunit Pol III is the most complex of the three eukaryotic polymerases. In addition to a ten-subunit enzymatic core and the two-subunit peripheral stalk, which are conserved in both Pol I and Pol II, Pol III comprises five additional subunits, which are organized in subcomplexes. The C82–C34–C31 heterotrimeric subcomplex is essential for TFIIIB-dependent recruitment and for promoter opening[17], whereas the C53–C37 heterodimeric subcomplex is involved in both initiation and termination of transcription[18,19].

Here we report the cryo-EM structures of an open Pol III PIC (OC-PIC) at 4.0 Å resolution, in which the DNA promoter has been spontaneously opened and the template strand engaged in the active site to form a full transcription bubble; a Pol III open complex (OC-POL3) at 3.4 Å resolution, which is similar to the OC-PIC but in which the upstream edge of the transcription bubble and TFIIIB are not resolved in the EM map; and unbound apo-Pol III (POL3) at 3.1 Å. Our results provide detailed insight into the mechanisms required for the initial steps of the Pol III transcription cycle, including DNA strand separation and template-strand engagement in the active site.

Visualization of the Pol III PIC

To obtain structural insights into initiation of Pol III transcription, we used size-exclusion chromatography to purify uncrosslinked Pol III PICs, assembled with a Brf1–TBP fusion protein[20] and with a fully complementary DNA scaffold grafted from the *SNR6* promoter (U6 snRNA) (Methods and Fig. 1a).

Using a wild-type Pol III PIC, we obtained a reconstruction of the OC-PIC at an overall resolution of 4.0 Å (40,887 particles, 18.2% of the total) and a reconstruction of the OC-POL3 at an overall resolution of 3.7 Å (OC1-POL3; 101,484 particles, 47.4% of the total) (Fig. 1b and Extended Data Figs 1a, b, 2a–c, 3a, b). For the Pol III PIC assembled in the presence of Bdp1Δ(355–372), a Bdp1 mutant that is deficient in promoter opening[21], we obtained another OC-POL3 (OC2-POL3; 100,237 particles, 21.5% of the total) at an overall resolution of 3.4 Å and an unbound POL3 (178,779 particles, 38.3% of the total) at 3.1 Å (Extended Data Figs 1c–e, 2d–f, 3c–e). We further sub-classified the latter into cPOL3 and oPOL3, according to the closed or open conformation of the C82–C34–C31 subcomplex, at an overall resolution of 3.3 Å and 3.4 Å, respectively (Extended Data Figs 1c–e, 3c–e). Because OC1-POL3 and OC2-POL3 are essentially identical and represent the same functional state, hereafter we refer only to OC2-POL3, which was solved at higher resolution, as OC-POL3.

Cryo-electron microscopy (cryo-EM) maps were used to build and refine atomic models of the Pol III PICs (Extended Data Table 1). The structure of the OC-PIC reveals that the spontaneously formed transcription bubble is tightly stabilized by both TFIIIB and Pol III subunits (Fig. 1b). The Pol III clamp is in a closed state, as observed in Pol II PICs[22,23]. Upon formation of the OC-PIC, regions of the specific Pol III subunits C37, C34 and C31, which were mobile in the unbound and elongating structures of Pol III, are now visible in the cryo-EM maps (Fig. 1b, Extended Data Fig. 4). The structure of OC-POL3 suggests that once the DNA is melted and the template strand loaded in the active site, Pol III retains the downstream edge of the bubble in a correct orientation, even when the main interactions between Pol III and TFIIIB are disrupted (Extended Data Fig. 1b, c), in agreement with previous studies[24]. This mechanism probably allows Pol III to maintain a correctly oriented bubble during promoter escape.

TFIIIB structure within the Pol III PIC

TFIIIB is centred around the TATA box, which is recognized specifically by TBP (Fig. 2a). The architecture of the N-terminal part of Brf1 (residues 1–264) closely resembles that of TFIIB in the Pol II PIC[22,23]. The Brf1 Zn-ribbon domain is inserted through the active site cleft

[1] The Institute of Cancer Research, London SW7 3RP, UK.

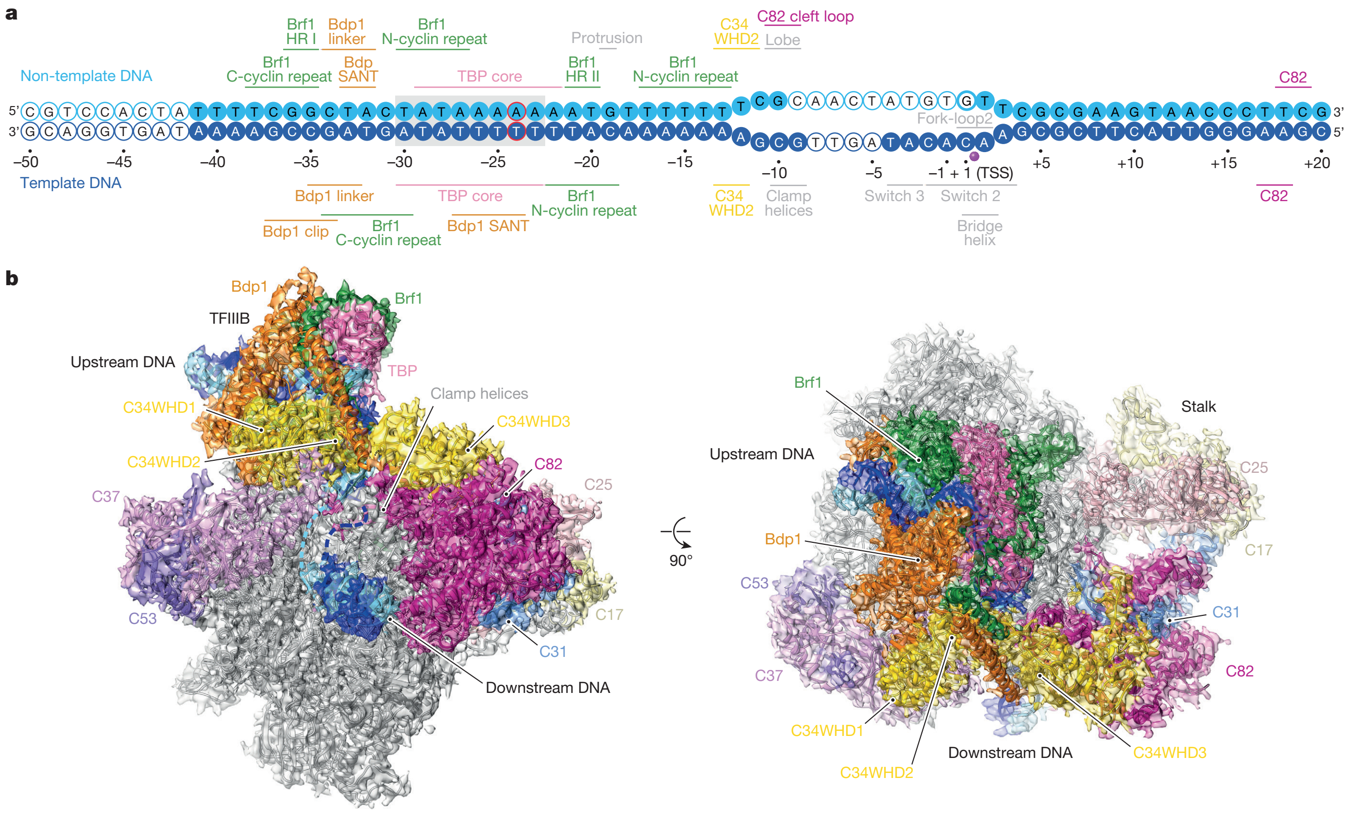


Figure 1 | Cryo-EM structure of the *Saccharomyces cerevisiae* RNA Pol III PIC. a, DNA nucleotides modelled in the OC-PIC structure are depicted as solid circles and numbered relative to the TSS of the *SRN6* gene. Mutation of nucleotide −24 to break the pseudo-symmetry of the TATA box is outlined in red. The position of the active site magnesium ion is indicated as a magenta sphere. The TATA box is highlighted with a grey box. Protein–DNA interactions are highlighted. **b**, Front and top view of yeast Pol III PIC structure (OC-PIC). Pol III core subunits are depicted in grey. Template and non-template strands of the DNA are shown in dark and light blue, respectively. Cryo-EM maps are shown as transparent surfaces. The same colour code is used throughout.

at the Pol III dock domain while the Brf1 C-terminal cyclin repeat is located between the Pol III wall and protrusion (Fig. 2a and Extended Data Fig. 4a, b). The linker region between the Brf1 Zn-ribbon and the N-terminal cyclin repeat, which in Pol II plays a role in stabilizing the template strand[22], is disordered in our structure. The Brf1 homology region I (HRI) includes a helical element, the 'Brf1 helical pin' that directly contacts the DNA, TBP and the Brf1 C-terminal cyclin repeat (Fig. 2a and Extended Data Fig. 4a, b). The Brf1 helical pin sequence PPSF/Y is strictly conserved in eukaryotes and is structurally homologous to the Brf2 molecular pin (amino acidic sequence PPCML); it stabilizes the Brf1–TBP–DNA complex, albeit without redox-sensing function[2] (Extended Data Fig. 5a, b). As previously described, the Brf1 homology region II (HRII) wraps around TBP and runs parallel to the promoter DNA[25] (Fig. 2a and Extended Data Fig. 4a, b). The Brf1 homology region III (HRIII), which forms a cryptic DNA binding domain[26], is not resolved in our structure.

Bdp1 adopts an intricate fold in the Pol III PIC (Fig. 2a and Extended Data Fig. 4a, c). The Bdp1 extended SANT domain and the Bdp1 linker bind to the major and minor grooves of the DNA, respectively, as observed in a human BDP1–BRF2–TBP–DNA complex[27]. At its N terminus, the Bdp1 linker extends towards the Pol III protrusion and forms a Bdp1 'tether' that interacts with the C37 'termination–initiation loop', which is disordered in cPOL3 and elongating Pol III[28] (Fig. 2b and Extended Data Fig. 4a, c, d). The Bdp1 tether and the C37 termination–initiation loop together form a composite 'Bdp1–C37 platform' (Fig. 2b) that aids the positioning and docking of the first and second winged helix domains (WHD1 and WHD2) of subunit C34 (Extended Data Fig. 4a, e). The C34 WHD1 and WHD2 are also stabilized only upon PIC formation (Fig. 2a, b), in agreement with previous photo-crosslinking studies[29,30]. The OC-PIC structure reveals that inactive Bdp1 deletion mutants that fail to open the promoter[21] map to the Bdp1 tether and linker regions. Thus, the common transcriptional defect of these mutants is attributable to the disruption of the correct topology and architecture of the Bdp1–C37 platform. In support of this, a Pol III enzyme that completely lacks subunit C37 displays a similar transcription initiation defect[18] and mutations of charged residues in the C34 WHD2 impair binding to TFIIIB and promoter opening[17].

N-terminally to the Bdp1 tether, two α-helices form the 'Bdp1 clip', a scissor-like element that binds the DNA in the major groove and contacts the Brf1 C-terminal cyclin repeat and the Bdp1 SANT domain (Fig. 2a, b and Extended Data Fig. 4a, c). At the C terminus of the extended Bdp1 SANT domain, Bdp1 folds into a long α-helix, the 'Bdp1 stem', that participates, with the Brf1 HRII and the Bdp1–C37 platform, in the rigid positioning of the C34 WHD2 over the cleft (Fig. 2a, b and Extended Data Fig. 4a, c, e).

TFIIIB forms an extensive number of contacts with the DNA and completely encloses the DNA promoter, explaining its unusually long half-life[27,31]. Upstream and downstream regions flanking the TATA box are bridged by TFIIIB, which acts as a molecular wrench that pivots around the C34 WHD2 and the Pol III protrusion.

DNA opening and template strand loading

In Pol III PICs, the formation of a transcription bubble occurs in a non-coordinated manner, with strand separation nucleating at the upstream edge of the transcription bubble and subsequently propagating downstream to the transcriptional start site (TSS)[24]. In our structure of the OC-PIC, the upstream edge of the transcription bubble is at nucleobase -12, in close agreement with biochemical data[24] and with the structure of the human open Pol II PIC[22] (Fig. 1a, b). The

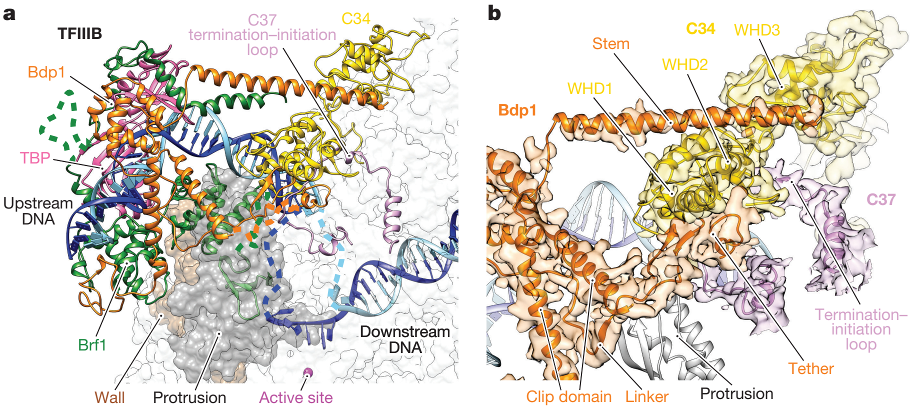


Figure 2 | Architecture of TFIIIB subunits and upstream DNA assembly. a, Positions of TFIIIB transcription factor and Pol III interacting subunits. Brf1, TBP, Bdp1, C37 and C34 are depicted as ribbon models; Pol III protrusion and wall domains are depicted as molecular surfaces. **b**, Detailed view of the Bdp1–C37 platform. Bdp1, C37 and C34 subunits are depicted as ribbon models in the cryo-EM map.

template and non-template strands are uncoupled by direct contact with Pol III subunits (Fig. 3a, b). The unwound template strand is in close proximity to an exposed hydrophobic pocket on the clamp helices, which we named the 'template strand pocket' (Fig. 3a). The template strand pocket is formed by Pol III residues W294, L298 and Y318, which are highly conserved, or substituted with chemically equivalent amino acids, across the eukaryotic kingdom (Fig. 3a and Extended Data Fig. 5c). The nucleobase at position −9 establishes Van der Waals interactions with W294, which stabilize the unwound DNA (Fig. 3a). The template strand pocket is specific for Pol III, as these residues are not conserved in Pol I and Pol II, except for L298 in Pol I (Extended Data Fig. 5c). Notably, part of the C160 'rudder' (Extended Data Fig. 4a), which is disordered in the OC-PIC, is stabilized in the structure of elongating Pol III[28], limiting the access of nucleic acids to the template strand pocket.

Concomitantly, the non-template strand is embraced between the C82 cleft loop and the Pol III C128 'lobe tip' (Fig. 3b and Extended Data Fig. 4a). Recognition and stabilization of the unwound non-template strand by the Pol III lobe tip also rationalizes the role of this region during transcription termination. Deletion mutants of the lobe tip display a termination read-through phenotype[32], suggesting a direct role for this region in stabilizing the terminator signal, a stretch of five or more thymidine residues on the non-template strand. Sensing of the terminator signal is likely to be carried out together with the C37 termination–initiation loop, another hot-spot of termination read-through mutants[33], which lies in close proximity to the lobe tip (Fig. 3b).

An extensive contact is formed between the DNA around positions −20 to −15 and a helix–turn–helix motif of Brf1, which is conserved in TFIIB[22] (Fig. 3a). In the opposite position to this site, the C34 WHD2 inserts a canonical recognition helix into the major groove of the DNA and the 'wing' at the site where the DNA is melted, stabilizing the unwound structure (Fig. 3a, b). The protrusion contacts the DNA backbone around register −19 and interacts with the C34 WHD2 recognition helix (Fig. 3a).

Once the DNA is unwound at the upstream edge of the transcription bubble, the template strand is engaged at the polymerase active site while the bubble is extended downstream. Comparison of the OC-PIC and POL3 structures reveals that, before PIC assembly, the path for loading the template is obstructed by a loop (residues 1,042–1,061 of subunit C128), which is part of the switch 3 region (Fig. 3c and Extended Data Fig. 4a). In the OC-PIC, the binding of the Brf1 Zn-ribbon reconfigures the switch 3 loop in an open conformation, as observed in elongating Pol II, which can directly stabilize the template strand (Fig. 3c). This structural rearrangement can explain the transcriptional defects observed for the Brf1 Zn-ribbon deletion mutant *in vitro* and *in vivo*[21]. Specifically, this Brf1 mutant is transcriptionally inactive owing to its inability to extend the upstream transcription

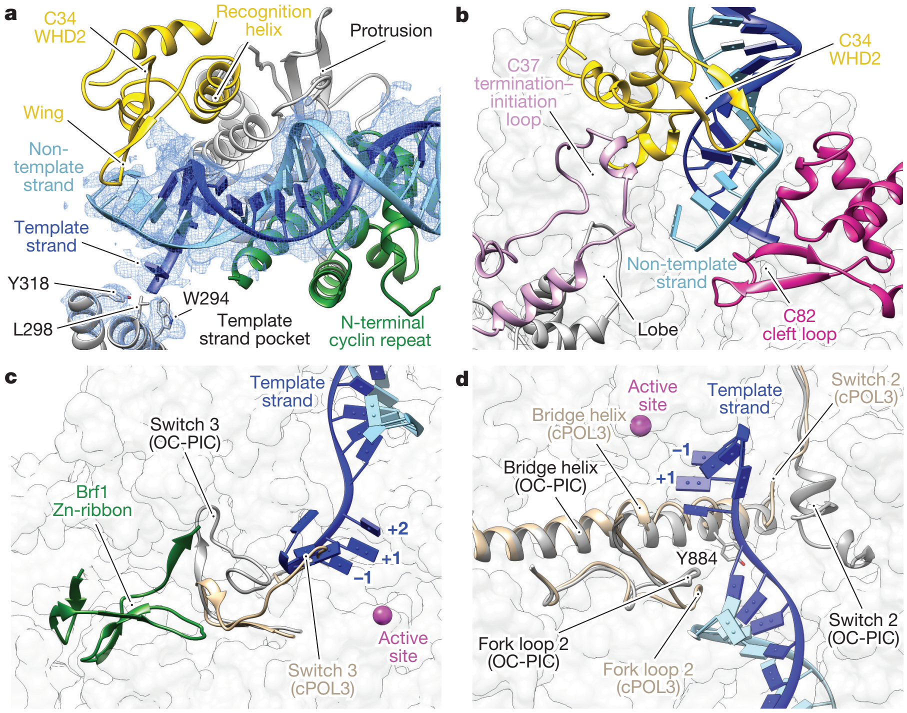


Figure 3 | Promoter opening and template strand loading. a, Detail of upstream promoter DNA opening. Side chains of residues involved in the template strand pocket are shown in stick representation. **b**, The non-template strand is threaded between the C82 cleft loop and C128 lobe tip. **c**, Detail of template DNA strand loading into the active site. The Brf1 Zn-ribbon domain reconfigures the Pol III switch loop 3 from a closed (cPOL3, wheat) to an open state (OC-PIC, grey). DNA nucleotides are numbered relative to the TSS. **d**, Stabilization of the DNA bubble downstream edge in the OC-PIC. Conserved Y884 residue is shown in stick representation.

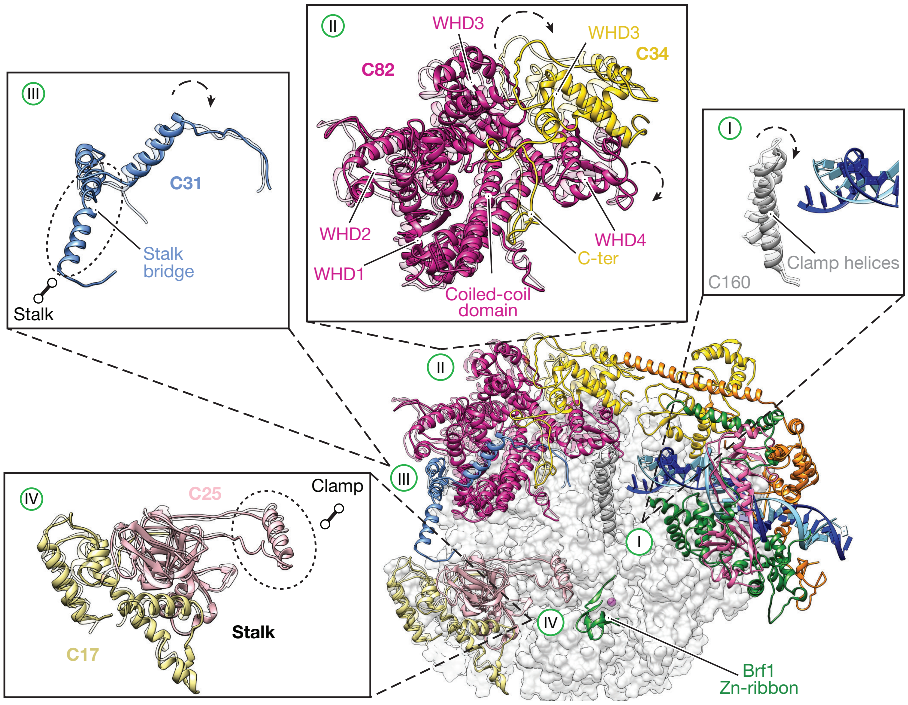


Figure 4 | Structural rearrangements upon open complex formation. Subunits involved in conformation changes are depicted as solid ribbon models. The core of Pol III is represented as a grey transparent surface. The numbering (I–IV) and close-up views refer to specific domain rearrangements occurring upon open PIC formation, which involve motions (dashed arrows), folding (dashed ellipsoids) and establishment of new contacts with neighbouring subunits (solid line with circles).

bubble, but its activity can be rescued by pre-opening the nucleic acid scaffolds at the downstream edge of the prospective bubble[21]. Thus, binding of the Brf1 Zn-ribbon induces a structural rearrangement that clears the cleft, facilitating loading of the template strand at the active site and formation of a full-size transcription bubble (Fig. 3c). The switch 3 loop adopts the same open conformation in the structure of the elongating Pol III[28], suggesting that the structural changes in the cleft prepare the OC-PIC for the elongation step.

Stabilization of the downstream edge of the bubble, which is in proximity to the active site, is reminiscent of that observed in the Pol II PICs[22] (Fig. 3d). The fork-loop 2 (Extended Data Fig. 4a) maintains the downstream transcription bubble by forming a wedge at the branching point, a universal function that seems to be conserved from bacterial to eukaryotic PICs[22]. The bridge helix, which in the unbound POL3 structure is kinked similarly to the paused or inhibited bacterial polymerase[34,35], is completely folded (Extended Data Figs 1, 4a) and interacts with the template strand through residue Y884, which is conserved in Pol I and Pol II (Fig. 3d and Extended Data Fig. 5d). The switch 2 is disordered in the POL3 structures but folds into a helical element in the OC-PIC and interacts with the template strand (Fig. 3d and Extended Data Fig. 4a). The structural changes responsible for the stabilization of the downstream edge of the transcription bubble are also observed in the structure of elongating Pol III[28], reinforcing the idea that the OC-PIC is primed for elongation.

Sensing the open complex formation

Opening of the promoter DNA is mediated by a cascade of structural changes involving several regions of Pol III and TFIIIB, raising the question of how promoter opening is detected and how the structural rearrangements are co-ordinated.

Upon TFIIIB binding, stabilization of the C34 WHD2 over the cleft and DNA melting at the upstream edge of the transcription bubble are accompanied by contraction of the clamp helices (Extended Data Fig. 4a, f), as compared to the POL3 or elongating Pol III structures[28] (Fig. 4, rearrangement I), to stabilize the unwound template strand. As a result, the C82–C34–C31 subcomplex, which lies on top of the clamp helices, is shifted towards the cleft (Fig. 4, rearrangement II) and undergoes a structural rearrangement that culminates in the stabilization of the C-terminal segment of subunit C34 (Fig. 4, rearrangement II and Extended Data Fig. 4a, e) and subunit C31 (Fig. 4, rearrangement III and Extended Data Fig. 4a, g), which are disordered in POL3 and the elongating Pol III[28] structures. The C31 'stalk bridge' folds into a α-helix that directly contacts the stalk, locking it at a defined angle with respect to the C82–C34–C31 subcomplex (Fig. 4, rearrangement III). The Pol III stalk is tightly anchored to the Pol III core and directly contacts the clamp through a region that is stabilized in our OC-PIC (Fig. 4, rearrangement IV). The Pol III stalk is mobile and can relay structural changes to the clamp[28], a mechanism shared with the archaeal and Pol II machineries[36,37]. Similarly, in these systems TFE and TFIIE directly bridge the stalk to the clamp through a Zn-ribbon domain[22,23,38].

Together, our data suggest that formation of the open PIC is sensed through concerted structural rearrangements that lock the Pol III clamp in a closed state, resulting in stabilization of the downstream edge of the transcription bubble (Fig. 4 and Supplementary Video S1). We observe the same arrangement of the C82–C34–C31 subcomplex, the stalk and the clamp in the OC-POL3 structure, suggesting that

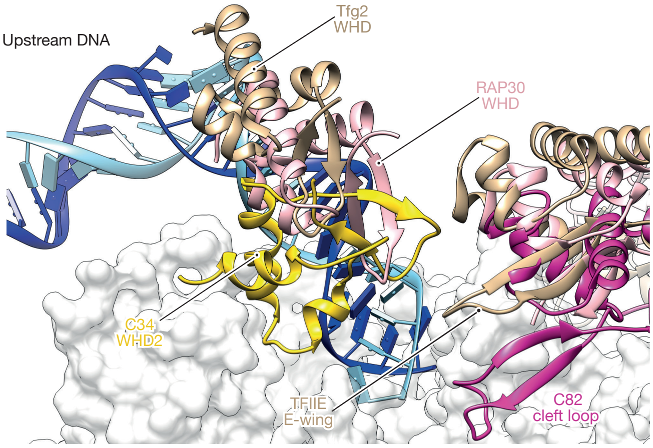


Figure 5 | The C82–C34–C31 subcomplex is a TFIIF–TFIIE hybrid. The upstream region of the DNA bubble is stabilized by the WHDs of subunit C34 in a similar manner to that of TFIIF subunits Tfg2 (wheat) and RAP30 (pink) in yeast and human, respectively. The non-template unwound DNA strand is stabilized by the C82 cleft loop, which resembles the yeast TFIIE E-wing element.

the locking mechanism of the clamp in a closed conformation allows Pol III to tightly hold the downstream edge of the transcription bubble (Extended Data Fig. 1b, c).

Comparison of eukaryotic PICs

The topology and architecture of the Pol III PIC closely resemble those of the Pol II PICs[22,23], whereas the Pol I PIC is more dissimilar as it has evolved and specialized for the transcription of a single gene[39–41] (Extended Data Fig. 6a). The relative orientation of the polymerase with respect to TBP is maintained owing to the architectural and functional conservation of Brf1 and TFIIB, which physically bridge the polymerase to TBP. The upstream and downstream edges of the transcription bubble are placed in the same locations (Extended Data Fig. 6a). A notable difference is observed in the upstream DNA pathway, as a result of the reduced TBP-induced bending of the DNA caused by the binding of the Bdp1 clip domain (Extended Data Fig. 6b).

The architecture of the protrusion displays remarkable differences among the ten-subunit cores of the three polymerases, which has implications for the assembly of the eukaryotic PICs. Whereas Pol I and Pol III use regions around the protrusion tip to directly contact the promoter DNA, albeit with Pol I doing so in a more extensive manner, the Pol II protrusion tip does not contact the DNA directly (Extended Data Fig. 6c). Pol III directly contacts the promoter through the conserved residue K409 of subunit C128.

In the Pol II PIC, the WHD of TFIIF Tfg2 (RAP30 in humans) binds and stabilizes the promoter DNA close to the upstream edge of the transcription bubble, in a manner which partially resembles C34 WHD2 (Fig. 5). However, whereas C34 WHD2 is directly involved in stabilizing the DNA at the site where strand separation occurs, the TFIIF Tfg2 WHD binds more upstream. From the opposite site, the TFIIE E-wing protrudes from the tip of the clamp helices towards the DNA. This structural element more strongly resembles the C82 cleft loop, which is, however, located more deeply in the cleft, stabilizing the non-template strand, while the E-wing has been proposed to play an important role in strand separation[23] (Fig. 5). Taking into account the functional conservation of TFIIE and C31 in physically bridging the stalk and the clamp, the C82–C34–C31 subcomplex can be regarded as a TFIIF–TFIIE hybrid rather than simply a TFIIE-like subcomplex[15]. The strong structural homology between the dimerization domain of the TFIIF-like C53–C37 subcomplex and TFIIF has been previously reported[28,42]. Comparison of the structures of the three eukaryotic PICs (Extended Data Fig. 6a) highlights the architectural and topological conservation between dissociable Pol II GTFs and stably associated Pol I and III specific subcomplexes, which have acquired specific functional roles during evolution[15,43].

Conclusions

The structures of the OC-PIC and OC-POL3 provide insights into the molecular basis of specific promoter recognition and opening. Our results converge with previous biochemical data[21,24] on a model that requires a concerted allosteric mechanism, involving TFIIIB and Pol III-specific subunits, to form a transcriptionally competent Pol III PIC (Extended Data Fig. 7, Fig. 4 and Supplementary Video S1). The structural data unveil how Pol III opens the promoter DNA in the absence of ATP hydrolysis, using solely the binding energy generated by the large number of newly established interactions formed during the assembly of an active PIC. Despite being structurally unrelated, the stabilization of unwound nucleobases by aromatic residues, trapping of the non-template strand and movements of the clamp are all aspects shared with the bacterial transcription machinery, which also unwinds DNA in an ATP-independent manner[44,45]. Archaeal[46], Pol I[39–41] and certain Pol II promoters[23] can be also opened without ATP hydrolysis, suggesting that DNA opening occurs through similar mechanisms across the three kingdoms of life.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 September; accepted 8 December 2017.

1. Filer, D. *et al.* RNA polymerase III limits longevity downstream of TORC1. *Nature* (2017).
2. Gouge, J. *et al.* Redox signaling by the RNA polymerase III TFIIB-related factor Brf2. *Cell* **163,** 1375–1387 (2015).
3. Goodarzi, H. *et al.* Modulated expression of specific tRNAs drives gene expression and cancer progression. *Cell* **165,** 1416–1427 (2016).
4. Dauwerse, J. G. *et al.* Mutations in genes encoding subunits of RNA polymerases I and III cause Treacher Collins syndrome. *Nat. Genet.* **43,** 20–22 (2011).
5. Thiffault, I. *et al.* Recessive mutations in POLR1C cause a leukodystrophy by impairing biogenesis of RNA polymerase III. *Nat. Commun.* **6,** 7623 (2015).
6. Lee, J., Moir, R. D. & Willis, I. M. Differential phosphorylation of RNA polymerase III and the initiation factor TFIIIB in *Saccharomyces cerevisiae*. *PLoS ONE* **10,** e0127225 (2015).
7. Fairley, J. A. *et al.* Direct regulation of tRNA and 5S rRNA gene transcription by Polo-like kinase 1. *Mol. Cell* **45,** 541–552 (2012).
8. Vannini, A. *et al.* Molecular basis of RNA polymerase III transcription repression by Maf1. *Cell* **143,** 59–70 (2010).
9. Harismendy, O. *et al.* Genome-wide location of yeast RNA polymerase III transcription machinery. *EMBO J.* **22,** 4738–4747 (2003).
10. Moqtaderi, Z. *et al.* Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. *Nat. Struct. Mol. Biol.* **17,** 635–640 (2010).
11. Barski, A. *et al.* Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nat. Struct. Mol. Biol.* **17,** 629–634 (2010).
12. Kassavetis, G. A., Letts, G. A. & Geiduschek, E. P. A minimal RNA polymerase III transcription system. *EMBO J.* **18,** 5042–5051 (1999).
13. Kassavetis, G. A., Braun, B. R., Nguyen, L. H. & Geiduschek, E. P. *S. cerevisiae* TFIIIB is the transcription initiation factor proper of RNA polymerase III, while TFIIIA and TFIIIC are assembly factors. *Cell* **60,** 235–245 (1990).
14. Dieci, G., Percudani, R., Giuliodori, S., Bottarelli, L. & Ottonello, S. TFIIIC-independent *in vitro* transcription of yeast tRNA genes. *J. Mol. Biol.* **299,** 601–613 (2000).
15. Vannini, A. & Cramer, P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Mol. Cell* **45,** 439–446 (2012).
16. Ishiguro, A., Kassavetis, G. A. & Geiduschek, E. P. Essential roles of Bdp1, a subunit of RNA polymerase III initiation factor TFIIIB, in transcription and tRNA processing. *Mol. Cell. Biol.* **22,** 3264–3275 (2002).
17. Brun, I., Sentenac, A. & Werner, M. Dual role of the C34 subunit of RNA polymerase III in transcription initiation. *EMBO J.* **16,** 5730–5741 (1997).
18. Kassavetis, G. A., Prakash, P. & Shim, E. The C53/C37 subcomplex of RNA polymerase III lies near the active site and participates in promoter opening. *J. Biol. Chem.* **285,** 2695–2706 (2010).
19. Arimbasseri, A. G. & Maraia, R. J. Mechanism of transcription termination by RNA polymerase III utilizes a non-template strand sequence-specific signal element. *Mol. Cell* **58,** 1124–1132 (2015).
20. Kassavetis, G. A., Soragni, E., Driscoll, R. & Geiduschek, E. P. Reconfiguring the connectivity of a multiprotein complex: fusions of yeast TATA-binding protein with Brf1, and the function of transcription factor IIIB. *Proc. Natl Acad. Sci. USA* **102,** 15406–15411 (2005).
21. Kassavetis, G. A., Letts, G. A. & Geiduschek, E. P. The RNA polymerase III transcription initiation factor TFIIIB participates in two steps of promoter opening. *EMBO J.* **20,** 2823–2834 (2001).
22. He, Y. *et al.* Near-atomic resolution visualization of human transcription promoter opening. *Nature* **533,** 359–365 (2016).
23. Plaschka, C. *et al.* Transcription initiation complex structures elucidate DNA opening. *Nature* **533,** 353–358 (2016).
24. Kassavetis, G. A., Blanco, J. A., Johnson, T. E. & Geiduschek, E. P. Formation of open and elongating transcription complexes by RNA polymerase III. *J. Mol. Biol.* **226,** 47–58 (1992).
25. Juo, Z. S., Kassavetis, G. A., Wang, J., Geiduschek, E. P. & Sigler, P. B. Crystal structure of a transcription factor IIIB core interface ternary complex. *Nature* **422,** 534–539 (2003).
26. Huet, J., Conesa, C., Carles, C. & Sentenac, A. A cryptic DNA binding domain at the COOH terminus of TFIIIB70 affects formation, stability, and function of preinitiation complexes. *J. Biol. Chem.* **272,** 18341–18349 (1997).
27. Gouge, J. *et al.* Molecular mechanisms of Bdp1 in TFIIIB assembly and RNA polymerase III transcription initiation. *Nat. Commun.* **8,** 130 (2017).
28. Hoffmann, N. A. *et al.* Molecular structures of unbound and transcribing RNA polymerase III. *Nature* **528,** 231–236 (2015).
29. Hu, H. L., Wu, C. C., Lee, J. C. & Chen, H. T. A region of Bdp1 necessary for transcription initiation that is located within the RNA polymerase III active site cleft. *Mol. Cell. Biol.* **35,** 2831–2840 (2015).
30. Wu, C. C., Lin, Y. C. & Chen, H. T. The TFIIF-like Rpc37/53 dimer lies at the center of a protein network to connect TFIIIC, Bdp1, and the RNA polymerase III active center. *Mol. Cell. Biol.* **31,** 2715–2728 (2011).
31. Cloutier, T. E., Librizzi, M. D., Mollah, A. K., Brenowitz, M. & Willis, I. M. Kinetic trapping of DNA by transcription factor IIIB. *Proc. Natl Acad. Sci. USA* **98,** 9581–9586 (2001).

32. Shaaban, S. A., Krupp, B. M. & Hall, B. D. Termination-altering mutations in the second-largest subunit of yeast RNA polymerase III. *Mol. Cell. Biol.* **15,** 1467–1478 (1995).
33. Rijal, K. & Maraia, R. J. RNA polymerase III mutants in TFIIFα-like C37 that cause terminator readthrough with no decrease in transcription output. *Nucleic Acids Res.* **41,** 139–155 (2013).
34. Weixlbaumer, A., Leon, K., Landick, R. & Darst, S. A. Structural basis of transcriptional pausing in bacteria. *Cell* **152,** 431–441 (2013).
35. Tagami, S. *et al.* Crystal structure of bacterial RNA polymerase bound with a transcription inhibitor protein. *Nature* **468,** 978–982 (2010).
36. He, Y., Fang, J., Taatjes, D. J. & Nogales, E. Structural visualization of key steps in human transcription initiation. *Nature* **495,** 481–486 (2013).
37. Schulz, S. *et al.* TFE and Spt4/5 open and close the RNA polymerase clamp during the transcription cycle. *Proc. Natl Acad. Sci. USA* **113,** E1816–E1825 (2016).
38. Jun, S. H. *et al.* The X-ray crystal structure of the euryarchaeal RNA polymerase in an open-clamp configuration. *Nat. Commun.* **5,** 5132 (2014).
39. Engel, C. *et al.* Structural basis of RNA polymerase I transcription initiation. *Cell* **169,** 120–131 (2017).
40. Han, Y. *et al.* Structural mechanism of ATP-independent transcription initiation by RNA polymerase I. *eLife* **6,** e27414 (2017).
41. Sadian, Y. *et al.* Structural insights into transcription initiation by yeast RNA polymerase I. *EMBO J.* **36,** 2698–2709 (2017).
42. Geiger, S. R. *et al.* RNA polymerase I contains a TFIIF-related DNA-binding subcomplex. *Mol. Cell* **39,** 583–594 (2010).
43. Carter, R. & Drouin, G. The increase in the number of subunits in eukaryotic RNA polymerase III relative to RNA polymerase II is due to the permanent recruitment of general transcription factors. *Mol. Biol. Evol.* **27,** 1035–1043 (2010).
44. Chakraborty, A. *et al.* Opening and closing of the bacterial RNA polymerase clamp. *Science* **337,** 591–595 (2012).
45. Zhang, Y. *et al.* Structural basis of transcription initiation. *Science* **338,** 1076–1080 (2012).
46. Hausner, W. & Thomm, M. Events during initiation of archaeal transcription: open complex formation and DNA-protein interactions. *J. Bacteriol.* **183,** 3025–3031 (2001).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank N. Cronin at the Institute of Cancer Research for help with yeast fermentation; C. Richardson for the computing infrastructure; C. Plaschka (MRC-LMB, Cambridge) for advice during data processing; and the staff at beamline M03 of the Diamond Light Source synchrotron (UK) for help with EM data collection (EM15629, EM16599 and EM166601). We acknowledge support and the use of resources of iNEXT, in particular C. Sachse and W. Hagen for EM data collection at EMBL Heidelberg (PID1956 and 2180). G.A.-P. is a recipient of a Marie Sklodowska-Curie Intra-European Fellowship (EU project 655238). E.M. is supported by Cancer Research UK (CR-UK C12209/A16749). A.V. is supported by a Biotechnology and Biological Sciences Research Council (BBSRC) New Investigator Award (BB/K014390/1), a Cancer Research UK Programme Foundation (CR-UK C47547/A21536) and a Wellcome Trust Investigator Award (200818/Z/16/Z).

Author Contributions G.A.-P. carried out yeast fermentation, Pol III purification, Pol III PIC reconstitution, EM specimen preparation, EM data collection and processing, model building and refinement. E.P.R. carried out EM data collection and processing of the Bdp1Δ(355–372) Pol III PIC and helped with EM data processing and analysis. F.B. carried out cryo-EM sample preparation, screening and sample collection. E.M. helped during initial EM characterization and data collection. A.V. designed and supervised research, analysed the structural data and prepared the manuscript with contributions from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to A.V. (Alessandro.Vannini@icr.ac.uk).

Reviewer Information *Nature* thanks R. Maraia and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Protein expression and purification. Endogenous Pol III, carrying a C-terminal Tap-tag on subunit AC40, was obtained through large-scale fermentation of a *Saccharomyces cerevisiae* SC1613 strain (Euroscarf) in a BioFlo Pro 75l Fermentor (New Brunswick Scientific). After overnight growth, the yeast culture was collected at an OD_{600} of ~10 and the cells were flash-frozen in liquid nitrogen. The sample was then subjected to lysis using a 6870D Freezer/Mill Cryogenic Grinder (4 cycles, 2 min run, 2 min cool, 15 cycles per second, SPEX Sample Prep). After cell lysis, the sample was resuspended in buffer A (50 mM Tris-HCl pH 8, 20% glycerol, 250 mM $(NH_4)_2SO_4$, 1 mM EDTA, 10 mM $MgCl_2$, 10 μM $ZnCl_2$, 12 mM BME, 1 mM PMSF and 2 mM benzamidine) and cleared by double centrifugation at 38,000*g* for 40 min. The soluble fraction was then filtered and loaded into a gravity column containing ~125 ml heparin sepharose 6 Fast Flow resin (GE Healthcare) equilibrated with buffer A. Next, the resin was washed with buffer B (buffer A + 0.5 mM EDTA, 1 mM $MgCl_2$, 10 μM $ZnCl_2$ and 1 mM BME in the absence of glycerol) and elution was performed in buffer C (buffer B + 1 M $(NH_4)_2SO_4$). Then, the sample was diluted to 500 mM $(NH_4)_2SO_4$ and incubated overnight at 4 °C with 10 ml IgG sepharose 6 Fast Flow resin (GE Healthcare). After washing, the beads were incubated with Tobacco Etch Virus protease (TEV) for 6 h at 4 °C. The TAP-tag cleaved sample was collected, diluted to ~60 mM $(NH_4)_2SO_4$ and loaded into a MonoQ 5/50 GL column (GE Healthcare). Elution was performed through a linear gradient from 60 mM $(NH_4)_2SO_4$ to 1 M $(NH_4)_2SO_4$ in 40 mM Tris-HCl pH 8, 0.5 mM EDTA, 1 mM $MgCl_2$, 10 μM $ZnCl_2$, 10 mM DTT, 1 mM PMSF and 2 mM benzamidine. Two peaks corresponding to RNA polymerase I and RNA polymerase III eluted at ~340 mM $(NH_4)_2SO_4$ and ~470 mM $(NH_4)_2SO_4$, respectively. The fractions corresponding to Pol III were collected, supplemented with 10% glycerol and stored at −80 °C.

In our reconstitution, we used a Brf1–TBP fusion protein, which has been shown to functionally substitute for Brf1 and TBP *in vitro* and *in vivo*[20]. Purification of the Brf1–TBP fusion protein was performed as previously described[20].

Wild-type Bdp1 or the transcriptionally inactive Bdp1Δ(355–372) mutant, which is unable to nucleate promoter DNA opening[21], were used for reconstitution of Pol III PICs. A pOPINJ plasmid containing *S. cerevisiae* Bdp1 genes was transformed into Rosetta (DE3) pLysS cells and grown in LSSB medium at 37 °C to an OD_{600} of ~0.7. The culture was then cooled at 4 °C for 1 h and induction was performed overnight with 1 mM IPTG at 15 °C. The collected cells were suspended in a buffer containing 10 mM Na_2HPO_4, 2 mM KH_2PO_4, 500 mM NaCl, 7 mM $MgCl_2$, 10 mM EDTA, 10% glycerol, 0.5% NP-40, 2.5 mM betaine, 10 mM BME, 0.5 mg/ml lysozyme, two protein inhibitor tablets (Roche) and a scoop of DNase I, and incubated at 4 °C for 30 min. The cell suspension was subjected to sonication (12 cycles, 15 s ON, 59 s OFF, 60% amplitude) and the lysate was fractionated by centrifugation at 48,000*g* and 4 °C for 40 min. Next, the soluble fraction was filtered and incubated for 2.5 h at 4 °C with ~5 ml glutathione sepharose 4 Fast Flow resin (GE Healthcare). Then, the beads were washed with 300 ml lysis buffer and the protein was eluted in 20 ml lysis buffer supplemented with 50 mM reduced glutathione. The collected sample was diluted to 100 mM NaCl with buffer QA (20 mM Tris-HCl pH 8.8, 0.2 mM EDTA, 10% glycerol and 1 mM DTT) and loaded into a HiTrap Q HP 5 ml column (GE Healthcare) equilibrated with 5% of buffer QB (buffer QA supplemented with 2 M NaCl). After washing, the sample was eluted through a linear gradient from 5% to 100% of buffer QB in 30 CV. The elution was analysed through SDS–PAGE and fractions containing GST–ScBdp1 were pooled and incubated overnight with 3C protease at 4 °C. The cleaved sample was concentrated to 5 ml and loaded into a HiLoad 16/600 Superdex 200-pg column equilibrated with 20 mM Tris-HCl pH 7.5, 150 mM NaCl and 2 mM DTT. The fractions corresponding to pure ScBdp1 were pooled, concentrated to ~2.5 mg/ml, flash-frozen in liquid nitrogen and stored at −80°C.

DNA oligonucleotides. The assembly of the Pol III PIC was performed using 70-nt complementary oligonucleotides based on the yeast *SRN6* gene promoter (template strand: CGAAGGGTTACTTCGCGAACACATAGTTGCGAAAAAA ACATTTTTTTATAGTAGCCGAAAATAGTGGACG and non-template strand: CGTCCACTATTTTCGGCTACTATAAAAAAATGTTTTTTTCGCAACTATGT GTTCGCGAAGTAACCCTTCG; Integrated DNA Technologies). Because of the pseudo-symmetrical nature of the U6 snRNA TATA box, which drives transcription in both directions in the absence of TFIIIC[12], we mutated a single nucleotide of the TATA box at position −24 from the TSS in order to favour unidirectional positioning of TFIIIB and avoid heterogeneity, as described in the TBP–Brf1(437–506) –DNA crystal structure[25] (Fig. 1a). The oligonucleotides were suspended in annealing buffer (50 mM Tris-HCl pH 8, 1 mM EDTA and 5 mM $MgCl_2$) to a final concentration of ~1 μM, mixed in equimolar concentration and heated at 95 °C for 5 min. In order to obtain dsDNA, the sample was cooled down to 10 °C at a rate of 1 °C per min.

Pol III PIC assembly. The formation of the Pol III PIC was carried out using 200 μg Pol III. First, the Brf1–TBP fusion protein and the annealed DNA were mixed at a 5:2.5 molar ratio (relative to Pol III) and incubated at room temperature for 30 min. Next, a fivefold excess of Bdp1 was added to the mixture and subjected to a similar incubation process. Then, Pol III was added and the mixture was diluted 12× with buffer containing 50 mM Tris pH 8, 10% glycerol, 3 mM DTT and 10 mM $MgCl_2$, to reduce the salt concentration. The sample was concentrated to ~300 μl with a Vivaspin 6 50K MWCO spin concentrator and incubated at room temperature for 30 min. Finally, the sample was applied to a Superose 6 10/300 GL column (GE Healthcare) equilibrated with 40 mM Tris-HCl pH 8, 80 mM NaCl, 3 mM DTT and 10 mM $MgCl_2$ and collected in 100-μl fractions. Analysis of the elution peaks by SDS–PAGE and silver staining indicated the presence of the Pol III and the transcription factors in the first peak. The fractions corresponding to the front of the peak were collected, concentrated to ~0.1 mg/ml and used in the subsequent EM analysis. In absence of Bdp1, a Pol III PIC could not be assembled using the conditions specified above.

Cryo-EM data collection. Cryo-EM samples were prepared in Quantifoil R 1.2/1.3 copper grids for wild-type Bdp1 (WT-PIC) sample or Quantifoil R2/2 Molybdenum grids for the Bdp1Δ(355–372) '(Bdp1Δ-PIC)' sample coated with a thin carbon film, which was made in house. The grids were glow discharged for 20 s at 15 mA using a PELCO EasyGlow glow discharger before sample addition. Sample (2 μl) was added to the grids and incubated for 30 s at 18 °C and 100% humidity. Then, the grid was blotted (drain time: 0.5 s, blot force: 13, blot time: 4–5 s) and plunge-frozen into liquid ethane using a VitroBot Mark IV (FEI) system.

Data were collected on a Titan Krios (FEI) transmission electron microscope at 300 keV using a Gatan Quantum energy filter and a K2 Summit direct detector. For the WT-PIC dataset, 4731 movies were collected at a 1.06 Å calibrated pixel size and a rate of 6 frames per second. Twenty-four frames were collected per movie within a defocus range from −1.6 μm to −3.4 μm and at a dose rate of 6.5 $e^-/Å^2$/s, which provided a total accumulated dose of 39 $e^-/Å^2$ (Extended Data Fig. 2a, Extended Data Table 1). For the Bdp1Δ-PIC dataset, a total of 6,220 micrographs were collected in super-resolution mode (0.5269 Å pixel size) at a rate of 1.25 frames per second. Twenty frames were collected for each movie within a similar defocus range to the wild-type dataset and at a dose rate of 2.52 $e^-/Å^2$/s, accumulating a total of ~40 $e^-/Å^2$ (Extended Data Fig. 2d and Extended Data Table 1).

Cryo-EM and image processing. MotionCor2[47] was used to perform the frame alignment and dose-weighting steps and CTFFIND 4.1.5[48] provided the estimation of the contrast transfer function (CTF) parameters. After CTF correction and movie alignment, approximately 10,000 particles were manually picked; 2D classes were calculated and used as references for automatic picking. All the subsequent steps of particle picking, extraction, classification and post-processing of refined models were performed in Relion 2.0.2[49].

For the WT-PIC dataset, ~836,000 particles were autopicked and selected for further 2D classification steps (Extended Data Fig. 2b). After three steps of particle classification, we obtained a dataset containing ~214,000 particles that was subjected to 3D classification using a 60 Å-filtered map of the apo-Pol III as a reference (Electron Microscopy Data Bank code: 3179, Extended Data Fig. 2c). This process provided four major classes. Class 2 (~91,000 particles) showed density corresponding to Pol III and the downstream open DNA but lacked density corresponding to TFIIIB or the upstream DNA. Class 3 (~34,000 particles) showed clear density for all the components and Class 4 (~59,000 particles) showed density of Pol III, the DNA and TFIIIB but at a lower threshold. Then, particles from Class 3 and Class 4 were joined and subjected to a second round of 3D classification (Extended Data Fig. 2c). This process provided a single class (Class 1, ~50,000 particles) corresponding to a Pol III–open DNA complex and another class (Class 2, ~43,000 particles) corresponding to a PIC (Pol III–DNA–TFIIIB at a similar density threshold). The PIC particles were refined and further subjected to polishing, producing a final map at 4.1 Å resolution according to the gold-standard FSC cut-off criterion at 0.143 (Extended Data Fig. 2c). In order to improve the quality in specific regions of the map, we performed focused 3D classifications without alignment using masks around TFIIIB–C34-tWHD and C82–C34–C31–stalk. This hierarchical process provided a map at 4.0 Å (OC-PIC) according to the 0.143 cut-off criterion (Extended Data Figs 2c, 3a). The results were validated by *ab initio* classification using Cryosparc[50]. Regarding the Pol III–DNA complexes, we joined the particles from Class 1 and Class 2 from the first round of 3D classification and Class 1 from the second round and performed a new classification process (Extended Data Fig. 2c). A predominant class (Class 1, ~101,000 particles) was obtained, which after data processing provided a map of Pol III–DNA at 3.7 Å (OC1-POL3) (Extended Data Fig. 3b).

The Bdp1Δ-PIC dataset was subjected to an initial 2D classification similar to that described for the WT-PIC data (Extended Data Fig. 2e). The resulting ~467,000 particles were 3D classified, which provided 3 classes (Classes 1, 3 and 4) corresponding to Pol III complexes with no TFIIIB visible (Extended Data Fig. 2f). These classes (~298,000 particles) were selected, joined and classified using a focused mask around the Pol III cleft and the downstream DNA (Extended Data Fig. 2f). This process provided 3 classes corresponding to an unbound Pol III (Classes 2, 3 and 4, ~178,000 particles) and a class of Pol III–DNA (Class 5, ~100,000 particles). The latter was subjected to 3D refinement and post-processing, which gave rise to a Pol III–DNA map (OC2-POL3) at 3.4 Å according to the 0.143 cut-off criterion (Extended Data Fig. 3c). For the unbound Pol III, we joined classes 2, 3 and 4 and subjected them to 3D refinement and post-processing, which gave rise to an unbound POLIII map (POL3) at 3.1 Å. Then, we performed a focused classification using a mask around the C82–C34–C31 subcomplex (Extended Data Fig. 2f). Two major classes (Class 1, ~62,000 particles and class 4, ~54,000 particles), representing both open and close states of the clamp helices and C82–C34–C31, were identified. These classes were refined and post-processed to 3.4 Å (oPOL3) and 3.3 Å resolution (cPOL3), respectively (Extended Data Fig. 3d, e). We noticed dissociation of Pol III from TFIIIB–DNA scaffolds only in the presence of the Bdp1Δ(355–372) mutant, suggesting a lower stability of the Pol III PIC and justifying the appearance of a large fraction of unbound apo-Pol III in the imaged EM samples. Indeed, in the absence of Bdp1, a Pol III PIC could not be assembled and purified by size-exclusion chromatography under the conditions used in this study. A substantial fraction of Pol III PICs assembled in the presence of Bdp1Δ(355–372) was still capable of efficiently melting the DNA (OC2-POL3, Extended Data Fig. 1c). This is likely to be due to the enhanced ability of the Brf1–TBP fusion protein to drive low levels of accurate transcription even in the absence of Bdp1, in contrast to isolated TBP and Brf1[20]. The OC1- and OC2-POL3 represent a state in which Pol III holds on the downstream edge of the correctly formed transcription bubble, even in the absence of TFIIIB, which is disordered or dissociated after formation of the open complex.

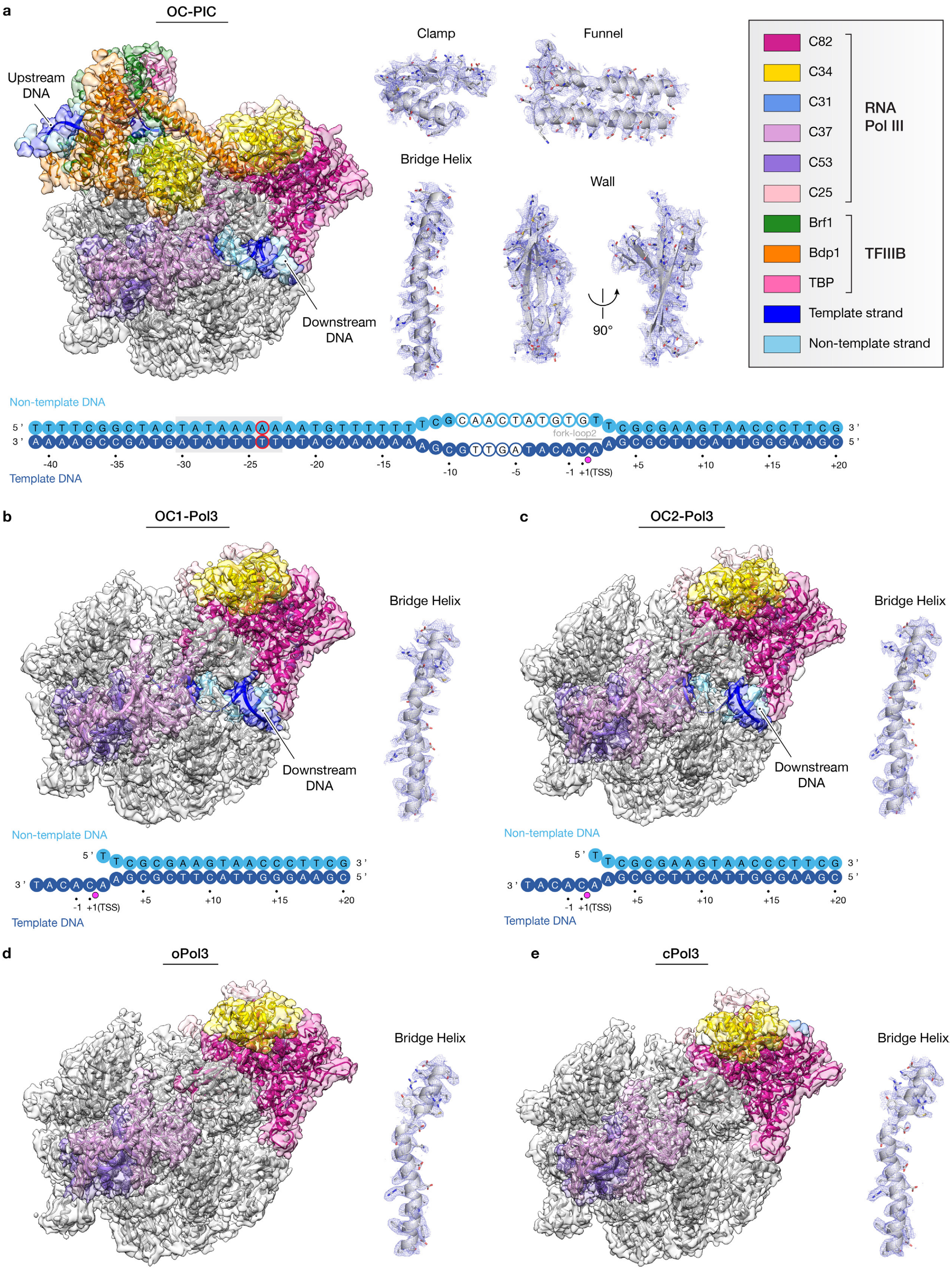
Model building and refinement. In an initial step, we fitted the model of Pol III elongation form (RCSB Protein Data Bank (PDB) code: 5FJ8) into our OC-PIC map, which confirmed the presence of all Pol III subunits and hinted at the existence of extra components such as TFIIIB or the C34 tandem winged-helix domain. Then, we fitted the models of Pol II open complexes (PDB codes: 5IYB and 5FYW) into the map, which strongly indicated the presence of the yeast Brf1–TBP complex in an equivalent position to TFIIB–TBP. Using the crystal structure of yeast TBP–Brf1–Cter (PDB code: 1NGM) and homology models of the Brf1 Zn-ribbon domain and cyclin folds generated with I-TASSER[51], we obtained an initial model of the yeast TFIIIB. Then, we identified the position of the Bpd1 SANT domain by fitting the human BRF2–TBP–BDP1 crystallographic model (PDB code: 5N9G) into this region. A homology model of the yeast Bdp1 SANT domain was also generated with I-TASSER[51]. Density corresponding to the Bdp1 linker was observed in the major groove of the DNA and it was manually extended using COOT[52]. This approach left unassigned one major density close to the Pol III protrusion and three helical densities protruding from the SANT domain. Considering previous studies[8,17], we inferred that the first density corresponded to C34 WH1 and WH2, which were absent in the previous Pol III models. We fitted the crystallographic models of the mouse homologue C39 (PDB codes: 2DK5 and 2DK8) and manually mutated and fitted in COOT[52]. Next, taking into consideration secondary structure predictions and crosslink data[29,30], we manually built the helical densities as part of Bdp1 N-terminal and C-terminal regions. Finally, in accordance with previous crosslinking results[29,30], the remaining small density close to C34 WHD1 and WHD2 was attributed to the Bdp1 tether region and to the termination–initiation loop of C37, which was flexible in previous structures. After preliminary positions for the different components had been assigned, we manually refined the models and built small unassigned regions using COOT[52]. The nucleic acids were initially placed using the human Pol II open complex (PDB code: 5IYB) as a guide and then manually fitted in the density using COOT[52]. Electron density supported the building of four nucleobases of the unwound template strand and three nucleobases of the unwound non-template strand, which are directly stabilized by Pol III subunits at the upstream edge of the transcription bubble, as well as seven nucleobases of the template strand, which are loaded into the active site, and two nucleobases of the non-template strand at the downstream edge of the bubble (Fig. 1a).

In the unbound Pol III (POL3, oPOL3, cPOL3), we observed two additional globular strong density peaks, as previously reported[28]. Thanks to the higher resolution of the reconstructions presented here, we confidently interpret these peaks as clusters of metal atoms, of unknown nature and function, stabilized in large mixed hydrophobic–hydrophilic pockets of Pol III (Extended Data Fig. 8). Because we are not sure about the nature of these metal atoms, we did not include them in the deposited models.

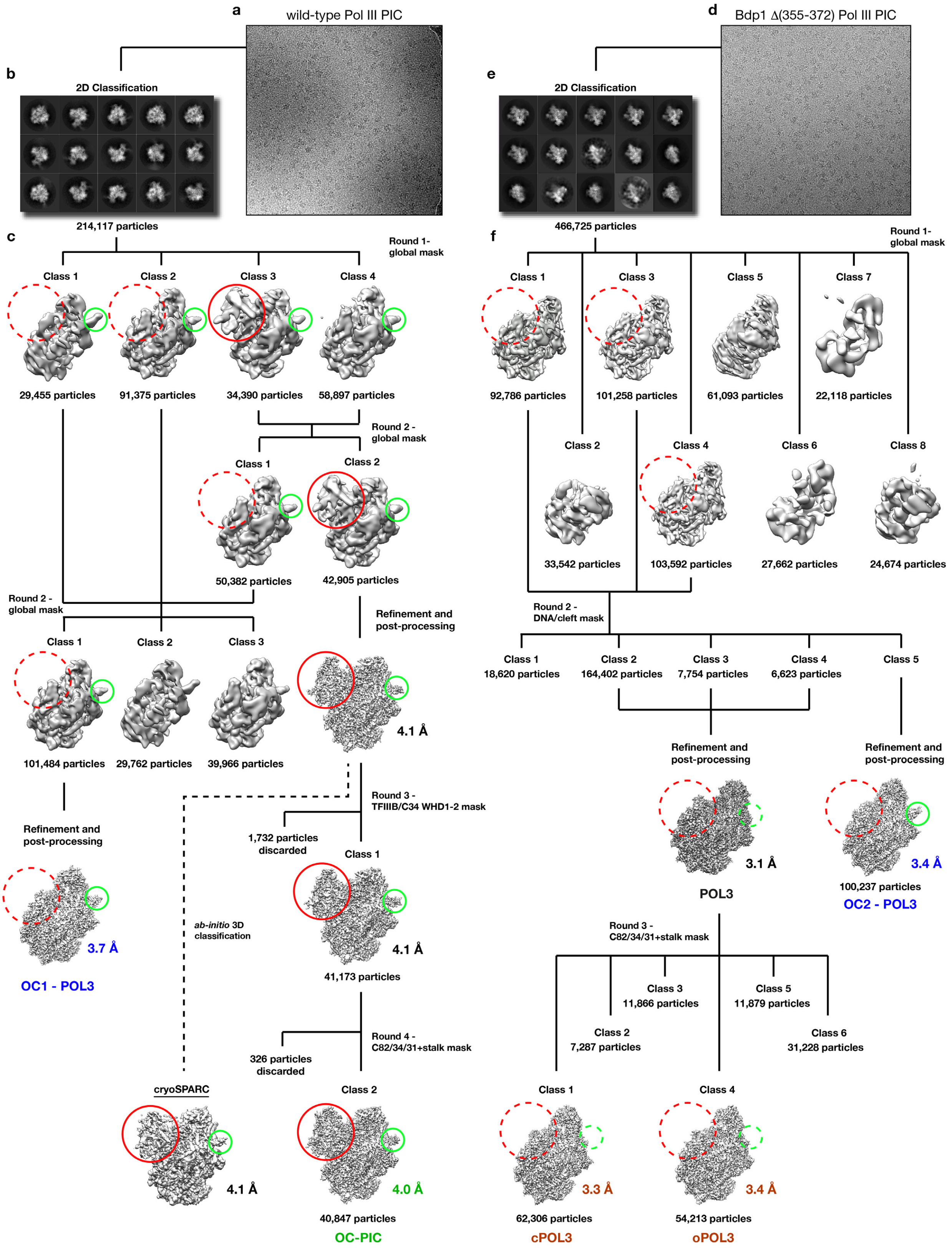
Refinement of the models against the maps was performed using Refmac5[53] and the PHENIX Suite[54]. Figures were prepared with Chimera UCSF[55] and Pymol (Schrödinger). Local resolution was calculated with ResMap 1.1.4[56] as implemented in Relion 2.0.2[49].

Data availability. Cryo-EM maps of OC-PIC, OC-POL3, POL3, cPOL3 and oPOL3 have been deposited in the Electron Microscopy Data Bank with accession codes EMD-3955 (OC-PIC), EMD-3956 (OC2-POL3), EMD-3959 (POL3), EMD-3957 (oPOL3) and EMD-3958 (cPOL3). The coordinates of the corresponding atomic models have been deposited in the Protein Data Bank under accession code 6EU0 (OC-PIC), 6EU1 (OC-POL3), 6EU2 (oPOL3) and 6EU3 (cPOL3).

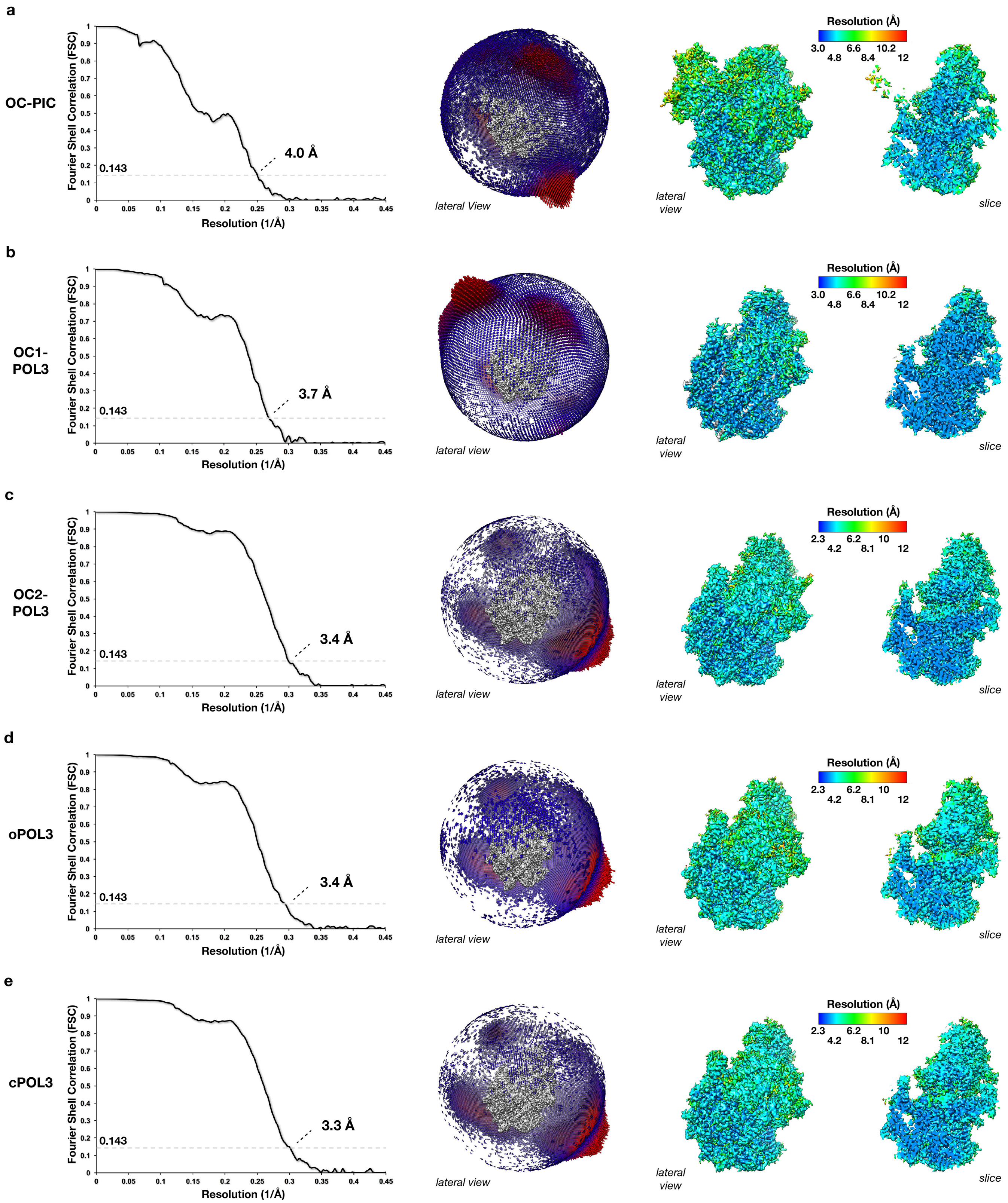
47. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14,** 331–332 (2017).
48. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192,** 216–221 (2015).
49. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5,** e18722 (2016).
50. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14,** 290–296 (2017).
51. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9,** 40 (2008).
52. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
53. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67,** 355–367 (2011).
54. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
55. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
56. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11,** 63–65 (2014).



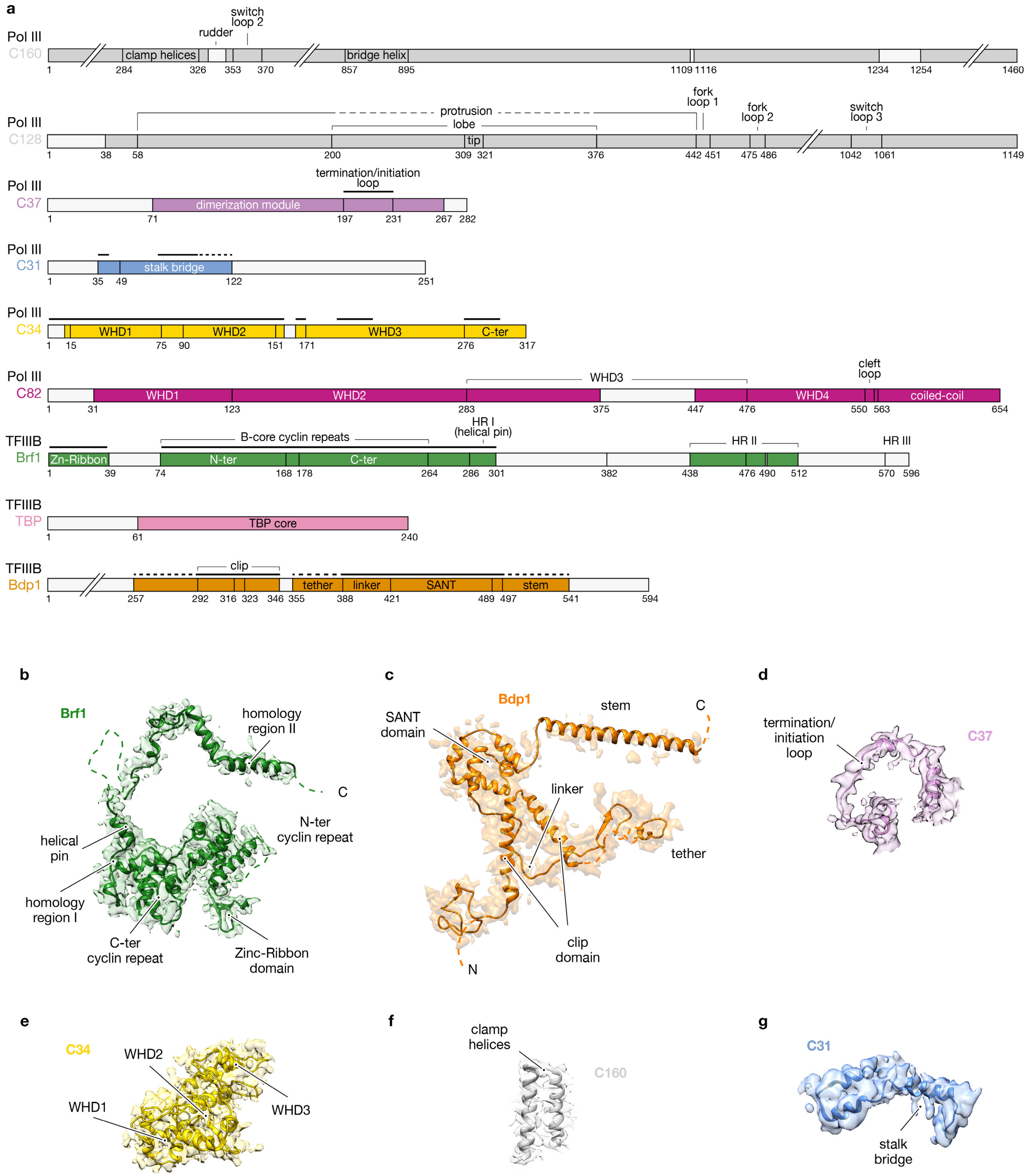
Extended Data Figure 1 | Cryo-EM reconstructions and model fitting. **a**, Cryo-EM reconstruction of the OC-PIC (left). Pol III core density is coloured in transparent grey and the TFIIIB, heterodimer, heterotrimer, stalk and DNA are coloured as indicated. Atomic models are represented as ribbon (right). Representative electron microscopy densities of different regions show the detail of the final reconstruction, where amino-acid side-chains are discernible as well as secondary structure features. Modelled DNA nucleotides are represented as indicated in Fig. 1. **b**, As in **a**, but for the OC1-POL3 reconstruction. **c**, As in **a**, but for the OC2-POL3 reconstruction. **d**, As in **a**, but for the oPOL3 reconstruction. **e**, As in **a**, but for the cPOL3 reconstruction.



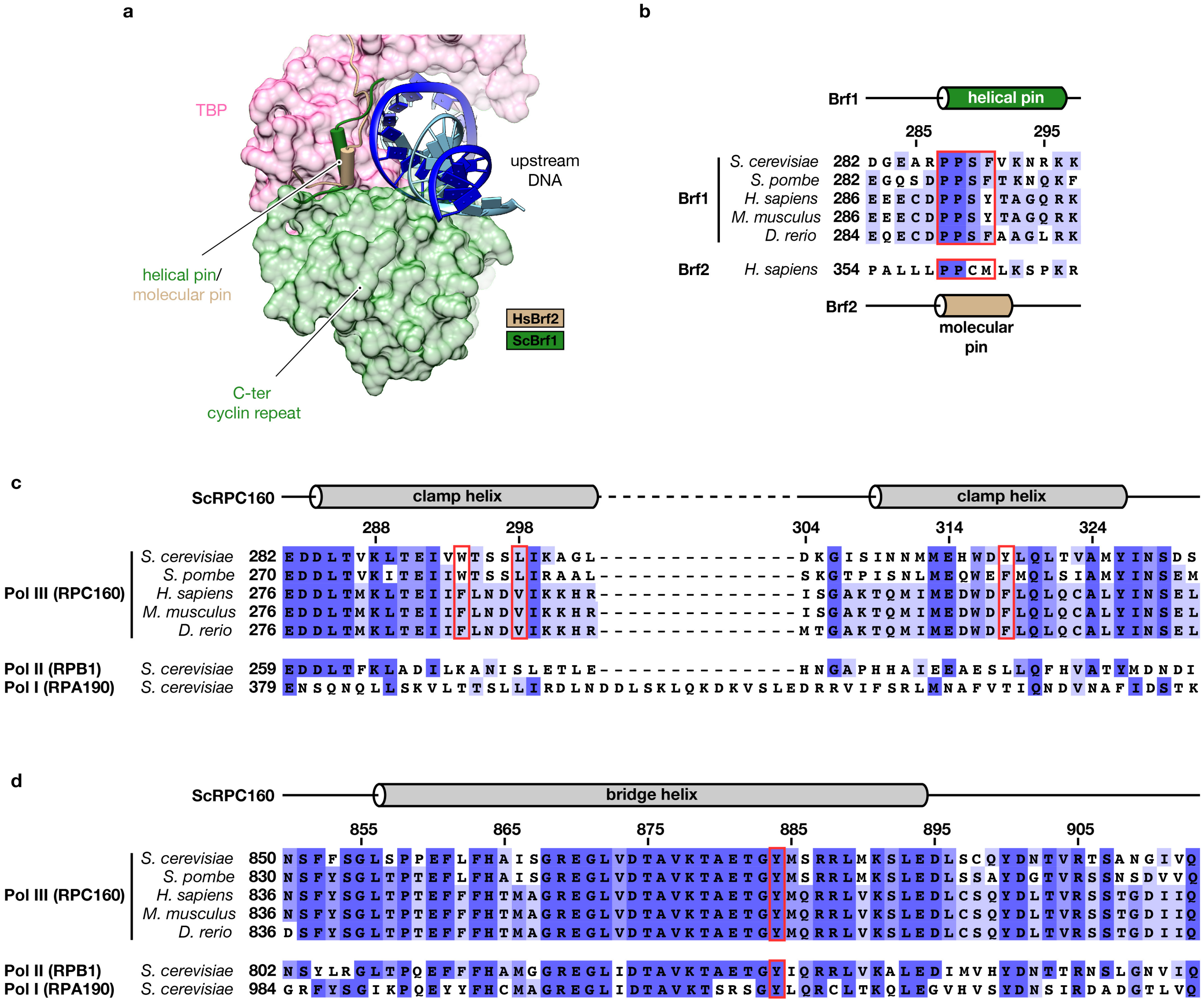
Extended Data Figure 2 | Cryo-EM data processing. a, **d**, Representative raw micrographs of the Bdp1 wild-type (**a**) and Bdp1(Δ355–372) (**d**) datasets. **b**, **e**, Fifteen representative reference-free 2D class averages of the wild-type Bdp1 (**b**) and Bdp1(Δ355–372) (**e**) datasets. **c**, 3D classification of the Bdp1 wild-type data set. The particles were subjected to a hierarchical process, which encompassed several rounds of classification using global masks or masks of specific regions of the complex, as indicated. The number of particles contributing to each class is indicated. The presence of densities corresponding to TFIIIB or the downstream DNA, which guided the classification process, is represented by red or green circles, respectively. **f**, As in **c**, but for the 3D classification of the Bdp1(Δ355–372) dataset.



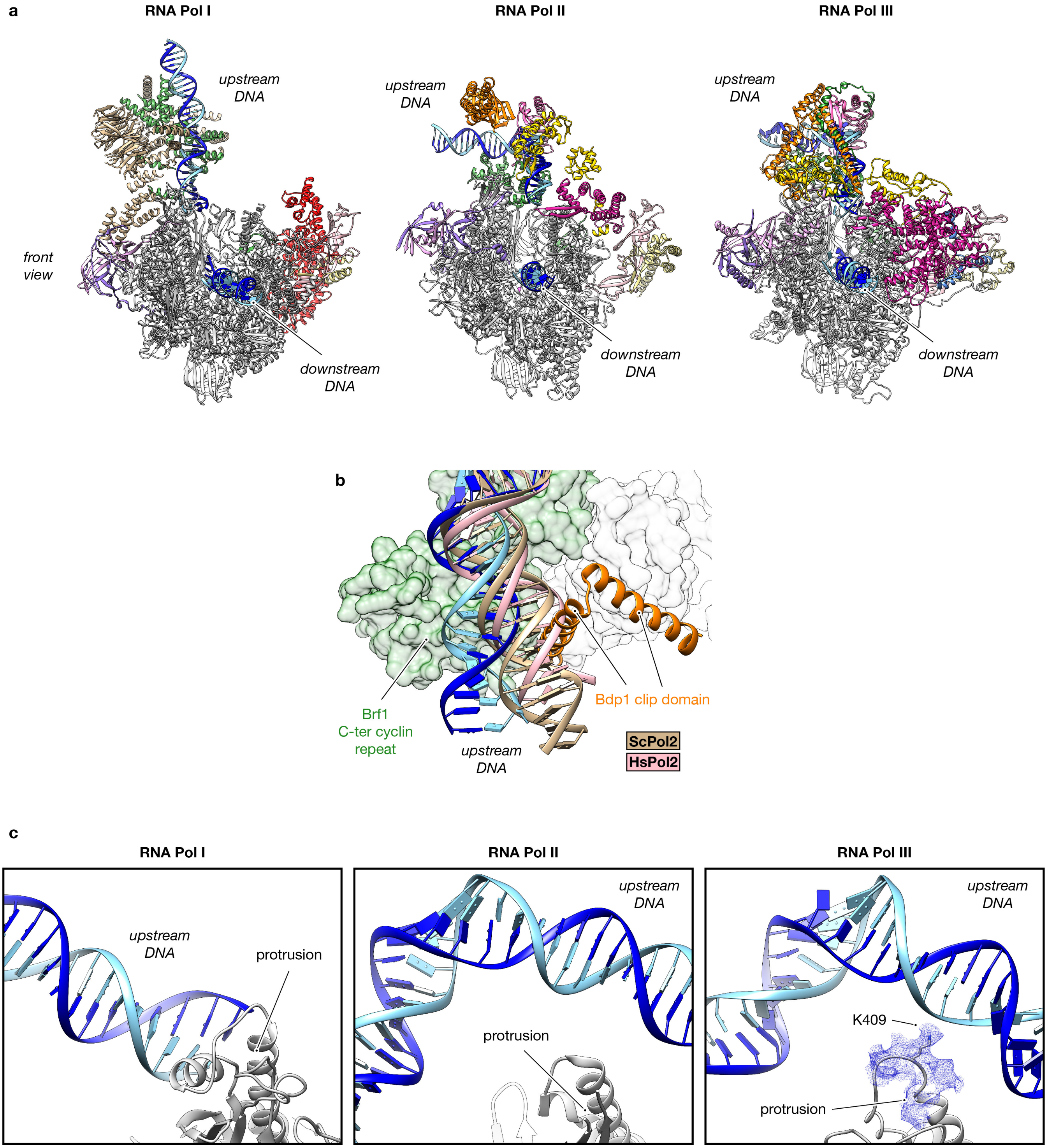
Extended Data Figure 3 | Resolution of cryo-EM reconstructions. **a**, Left, Fourier shell correlation plot of the OC-PIC reconstruction with the estimated resolution at the gold-standard FSC (FSC = 0.143). Middle, lateral view of the orientation distribution sphere of the particles that contributed to the OC-PIC reconstruction. The heights of the surface bars indicate the relative number of particles in a given orientation. Right, resolution estimation represented by a lateral view and central slice of the OC-PIC cryo-EM map The map is coloured according to the local resolution, as indicated in the scale bar. Local resolution was calculated with ResMap 1.1.4[56] as implemented in Relion 2.0.2[49]. **b**–**e**, As in **a**, but for the OC1-POL3 (**b**), OC2-POL3 (**c**), oPOL3 (**d**) and cPOL3 (**e**) reconstructions.



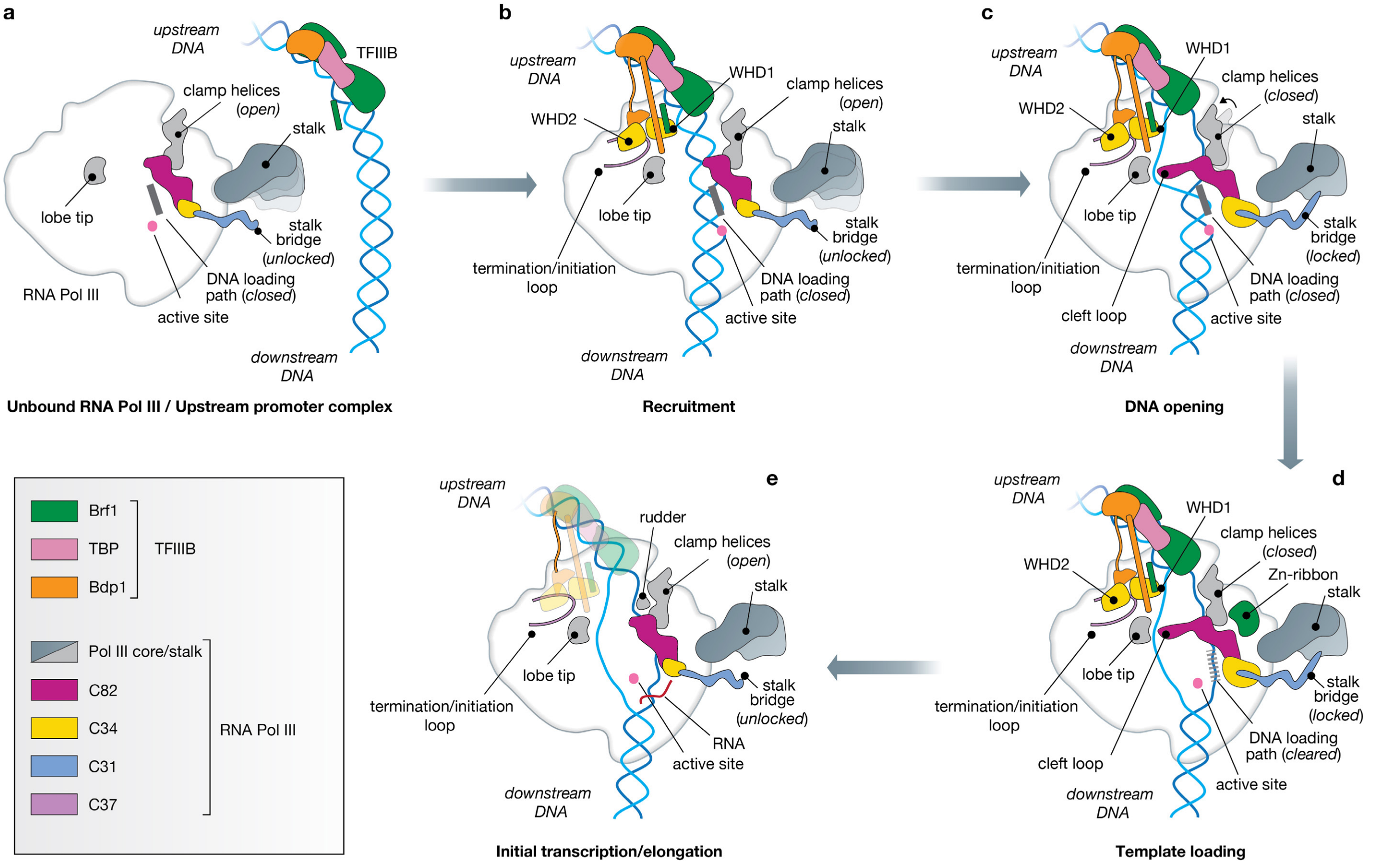
Extended Data Figure 4 | Architecture of TFIIIB and Pol III subunits involved in PIC assembly. a, Domain architecture of TFIIIB and Pol III subunits. Protein regions are depicted according to their presence (solid colour boxes) or absence (empty boxes) in the OC-PIC structure. Regions built *de novo* (for which previous structural information was not available) are highlighted with a black line (full atomic model) or with a dashed black line (backbone model). The same colour scheme is used for ribbon models and cryo-EM maps in **b**–**g**. **b**, *S. cerevisiae* Brf1 domain architecture. Regions absent in the density are indicated as a dashed line. C and N termini are indicated. **c**, Bdp1 domain architecture. Regions absent in the density are indicated as a dashed line. C and N termini are indicated. **d**, C37 termination–initiation loop architecture. **e**, C34 domain architecture. **f**, Clamp helices architecture. **g**, C31 stalk bridge architecture.


a
b
TBP
upstream DNA
helical pin/
molecular pin
C-ter cyclin repeat
HsBrf2
ScBrf1
Brf1
helical pin
285
295
Brf1
S. cerevisiae 282 DGEARPPSFVKNRKK
S. pombe 282 EGQSDPPSFTKNQKF
H. sapiens 286 EEECDPPSYTAGQRK
M. musculus 286 EEECDPPSYTAGQRK
D. rerio 284 EQECDPPSFAAGLRK
Brf2 H. sapiens 354 PALLLPPCMLKSPKR
Brf2
molecular pin
c
ScRPC160
clamp helix
clamp helix
288
298
304
314
324
Pol III (RPC160)
S. cerevisiae 282 EDDLTVKLTEIVWTSSLIKAGL---------------DKGISINNMMEHWDYLQLTVAMYINSDS
S. pombe 270 EDDLTVKITEIIWTSSLIRAAL---------------SKGTPISNLMEQWEFMQLSIAMYINSEM
H. sapiens 276 EDDLTMKLTEIIFLNDVIKKHR---------------ISGAKTQMIMEDWDFLQLQCALYINSEL
M. musculus 276 EDDLTMKLTEIIFLNDVIKKHR---------------ISGAKTQMIMEDWDFLQLQCALYINSEL
D. rerio 276 EDDLTMKLTEIIFLNDVIKKHR---------------MTGAKTQMIMEDWDFLQLQCALYINSEL
Pol II (RPB1) S. cerevisiae 259 EDDLTFKLADILKANISLETLE---------------HNGAPHHAIEEAESLLQFHVATYMDNDI
Pol I (RPA190) S. cerevisiae 379 ENSQNQLLSKVLTTSLLIRDLNDDLSKLQKDKVSLEDRRVIFSRLMNAFVTIQNDVNAFIDSTK
d
ScRPC160
bridge helix
855
865
875
885
895
905
Pol III (RPC160)
S. cerevisiae 850 NSFFSGLSPPEFLFHAISGREGLVDTAVKTAETGYMSRRLMKSLEDLSCQYDNTVRTSANGIVQ
S. pombe 830 NSFYSGLTPTEFLFHAISGREGLVDTAVKTAETGYMSRRLMKSLEDLSSAYDGTVRSSNSDVVQ
H. sapiens 836 NSFYSGLTPTEFFFHTMAGREGLVDTAVKTAETGYMQRRLVKSLEDLCSQYDLTVRSSTGDIIQ
M. musculus 836 NSFYSGLTPTEFFFHTMAGREGLVDTAVKTAETGYMQRRLVKSLEDLCSQYDLTVRSSTGDIIQ
D. rerio 836 DSFYSGLTPTEFFFHTMAGREGLVDTAVKTAETGYMQRRLVKSLEDLCSQYDLTVRSSTGDIIQ
Pol II (RPB1) S. cerevisiae 802 NSYLRGLTPQEFFFHAMGGREGLIDTAVKTAETGYIQRRLVKALEDIMVHYDNTTRNSLGNVIQ
Pol I (RPA190) S. cerevisiae 984 GRFYSGIKPQEYYFHCMAGREGLIDTAVKTSRSGYLQRCLTKQLEGVHVSYDNSIRDADGTLVQ

Extended Data Figure 5 | Structural conservation. a, Structure alignment of *S. cerevisiae* Brf1 and *H. sapiens* BRF2 (PDB code: 5N9G). *S. cerevisiae* TBP (pink) and Brf1 B-core cyclin repeats (green) are represented as molecular surfaces. Brf1 helical pin and Brf2 molecular pin are shown as green and wheat cylinders, respectively. **b**, Sequence alignment of *S. cerevisiae* Brf1 helical pin and *H. sapiens* BRF2 molecular pin. Residues are coloured according to their percentage identity, with dark and light blue indicating high and low sequence identity, respectively. Structurally conserved residues between the Brf1 helical pin and the Brf2 molecular pin are outlined in red. **c**, Multiple-sequence alignment of the clamp helices of RPA190 (Pol I), RPB1 (Pol II) and RPC160 (Pol III). Residues are coloured according to their percentage identity, with dark and light blue indicating high and low sequence identity, respectively. Residues participating in the template strand pocket (W294, L298 and Y318) are outlined in red. **d**, Multiple-sequence alignment of the bridge helices of RPA190 (Pol I), RPB1 (Pol II) and RPC160 (Pol III), coloured as in **c**. Conserved residue Y884 is outlined in red.



Extended Data Figure 6 | General comparison of Pol I, Pol II and Pol III PICs. a, Front views of *S. cerevisiae* Pol I PIC (PDB code: 5OA1), Pol II OC (PDB code: 5FYW) and Pol III PIC (this work). Pol III subunits are coloured as in Fig. 1. Colour scheme of Pol I and Pol II is based on architectural similarities to the Pol III system. **b**, Upstream DNA path differences. The DNA pathway in the Pol III PIC (light and dark blue) is different from that in yeast (wheat, PDB code: 5FYW) or human (pink, PDB code: 5IYB) Pol II PICs, probably owing to the interaction with the Bdp1 clip domain (orange). **c**, Comparison of Pol I, Pol II and Pol III protrusion tip in the PICs. Pol I (PDB code: 5OA1) and Pol III (this work) contact the promoter DNA through residues of the protrusion tip. Pol I participates in an extensive network of interactions that involve the binding of a α-helix to the major groove of the DNA, whereas Pol III binds to the non-template strand of the DNA (light blue) through the conserved residue K409. Pol II protrusion (PDB code: 5FYW) does not participate in direct contacts with the DNA.



Unbound RNA Pol III / Upstream promoter complex

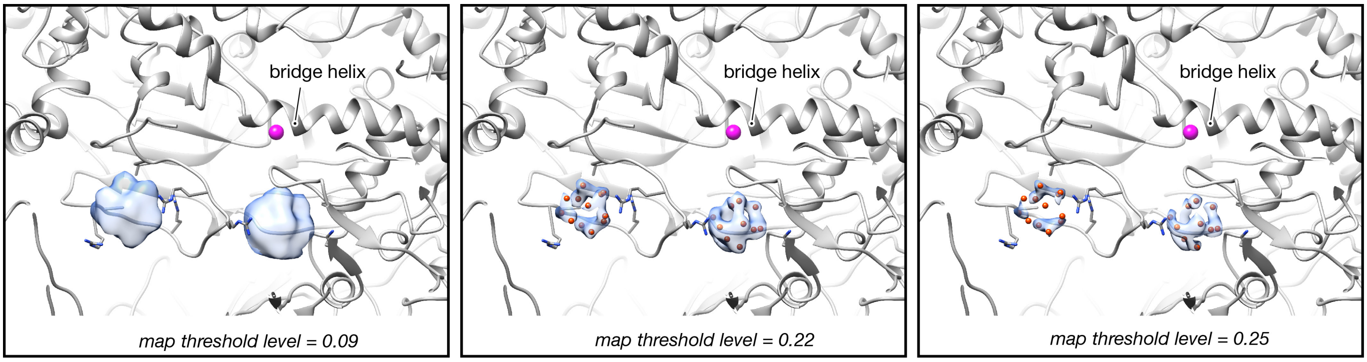
Recruitment

DNA opening

Initial transcription/elongation

Template loading

Extended Data Figure 7 | Mechanism of Pol III transcription initiation. **a**, In unbound Pol III the stalk region and clamp are mobile and can adopt an open or closed conformation. **b**, Upon TFIIIB recruitment, C34 WHD1 and WHD2 are positioned over the cleft through the interaction with the Bdp1 tether and the C37 termination–initiation loop. **c**, The C34 WHD2 and the Brf1 N-terminal cyclin fold promote DNA melting, which occurs through stabilization of the template strand in a template strand pocket of the clamp helices, and the non-template strand between the C82 cleft loop and the C128 tip lobe domain. Contraction of the clamp helices induces a conformational change in the C82–C34–C31 subcomplex. The stalk and clamp are now locked by the C31 stalk bridge. **d**, The template strand is loaded in the active site and the transcription bubble is fully expanded, as binding of the Brf1 Zn-ribbon domain clears the DNA loading pathway. The template strand is correctly engaged in the active site cleft, in a configuration primed for elongation. The transcription bubble around the active site is very stable and might even be maintained in circumstances in which the main contacts with TFIIIB are disrupted. The clamp is locked in a closed conformation that prevents the re-annealing of the transcription bubble. This might be particularly important during promoter escape, as short Pol III DNA–RNA hybrids are less tightly bound than in Pol II[28]. The stalk and clamp are locked in a closed state that prevents bubble reannealing. **e**, RNA synthesis starts, the clamp helices are released and the clamp and stalk are now unlocked. The clamp remains closed during elongation but can re-open during the following steps of the transcription cycle. The rudder is repositioned and occludes access to the template strand pocket, presumably to ensure Pol III processivity.


bridge helix
map threshold level = 0.09
bridge helix
map threshold level = 0.22
bridge helix
map threshold level = 0.25

Extended Data Figure 8 | High density peaks in unbound Pol III cryo-EM reconstructions. The globular density peaks (blue) observed in the cryo-EM maps of the unbound Pol III reconstructions are represented at three different threshold levels. The strong features of these regions are observed even at high threshold levels, suggesting the presence of metal clusters in hydrophilic–hydrophobic pockets of Pol III. Pol III core subunits are depicted in grey and the active site magnesium ion is represented as a magenta sphere.

Extended Data Table 1 | Cryo-EM data collection, refinement and model statistics

	OC-PIC (EMD-3955) (PDB 6EU0)	OC2-POL3 (EMD-3956) (PDB 6EU1)	oPOL3 (EMD-3957) (PDB 6EU2)	cPOL3 (EMD-3958) (PDB 6EU3)
Data collection and processing				
Voltage (kV)	300	300	300	300
Electron exposure ($e^{-}/Å^{2}$)	39	40	40	40
Defocus range (μm)	-1.6 to -3.4	-1.6 to -3.4	-1.6 to -3.4	-1.6 to -3.4
Pixel size (Å)	1.06	0.526	0.526	0.526
Reconstruction (RELION)				
Initial particle images (no.)	214,117	466,725	466,725	466,725
Final particle images (no.)	40,847	100,237	54,213	62,306
Map resolution (Å)	4.0	3.4	3.4	3.3
FSC threshold	(0.143-thr)	(0.143-thr)	(0.143-thr)	(0.143-thr)
Map sharpening *B* factor ($Å^{2}$)	101.26	88.99	78.34	83.31
Model composition				
Non-hydrogen atoms	48,919	40,602	38,329	38,330
Protein residues	5949	5019	4825	4825
Refinement (PHENIX)				
Map CC	0.655	0.683	0.737	0.734
R.m.s. deviations				
Bond lengths (Å)	0.00	0.02	0.03	0.02
Bond angles (°)	1.05	1.12	1.07	1.05
Validation				
MolProbity score	2.12	2.01	1.92	1.90
Clashscore (all-atom)	8.30	6.03	5.37	5.09
Poor rotamers (%)	0.49	0.32	0.35	0.28
Ramachandran plot				
Favored (%)	84.22	83.60	86.53	86.61
Allowed (%)	15.43	16.14	13.01	12.89
Disallowed (%)	0.36	0.26	0.46	0.50

doi:10.1038/nature24999

Black-hole-regulated star formation in massive galaxies

Ignacio Martín-Navarro[1,2], Jean P. Brodie[2], Aaron J. Romanowsky[1,3], Tomás Ruiz-Lara[4,5] & Glenn van de Ven[2,6]

Supermassive black holes, with masses more than a million times that of the Sun, seem to inhabit the centres of all massive galaxies[1,2]. Cosmologically motivated theories of galaxy formation require feedback from these supermassive black holes to regulate star formation[3]. In the absence of such feedback, state-of-the-art numerical simulations fail to reproduce the number density and properties of massive galaxies in the local Universe[4–6]. There is, however, no observational evidence of this strongly coupled coevolution between supermassive black holes and star formation, impeding our understanding of baryonic processes within galaxies. Here we report that the star formation histories of nearby massive galaxies, as measured from their integrated optical spectra, depend on the mass of the central supermassive black hole. Our results indicate that the black-hole mass scales with the gas cooling rate in the early Universe. The subsequent quenching of star formation takes place earlier and more efficiently in galaxies that host higher-mass central black holes. The observed relation between black-hole mass and star formation efficiency applies to all generations of stars formed throughout the life of a galaxy, revealing a continuous interplay between black-hole activity and baryon cooling.

As shown in Fig. 1, the mass of supermassive black holes ($M_\bullet$) scales with the stellar velocity dispersion (σ) of their host galaxies[2,7]. The scatter in this relation can be used to quantify how massive a given black hole is compared with the average population. We can then define over-massive and under-massive black-hole galaxies as those objects lying, respectively, above and below the best-fitting relation between $M_\bullet$ and σ. In other words, over-massive black-hole galaxies have central black holes more massive than expected, whereas under-massive black-hole galaxies host relatively light supermassive black holes. The distinction between these two types of galaxy allows us to evaluate the role of black-hole activity in star formation, as the amount of energy released into a galaxy is proportional to the mass of the black hole[8,9].

We based our stellar population analysis on long-slit optical spectra from the Hobby–Eberly Telescope Massive Galaxy Survey (HETMGS)[10]. The resolution of the data varies between 4.8 Å and 7.5 Å, depending on the slit width. We adopted a fixed aperture of half the effective radius, $0.5R_e$, where R_e is defined as the galactocentric radius that encloses half of the total light of a galaxy. This aperture is large enough to allow a direct comparison in the future between our results and numerical simulations, but also small enough to ensure that we are dominated by *in situ* star formation[11]. The sizes of all galaxies in our sample were calculated in a homogeneous way using infrared *K*-band photometry[12]. We focused on spectroscopic analysis of wavelengths between 460 and 550 nm, covering the most prominent spectral features indicating age and metallicity in the optical range.

Star formation histories (SFHs) were measured using the Stellar Content and Kinematics via Maximum A Posteriori likelihood (STECKMAP) code[13], fed with the MILES stellar population synthesis models[14]. STECKMAP is a Bayesian method that decomposes the observed spectrum of a galaxy as a temporal series of single stellar population models. Its ability to recover reliable SFHs of unresolved systems has been thoroughly tested[15–17]. Furthermore, STECKMAP-based SFHs are in remarkable agreement with those based on colour–magnitude diagrams of nearby resolved systems[18]. Our SFHs are reliable in a relative sense (see Methods) even if there are systematics from the limited set of models.

Our final sample consists of all HETMGS galaxies for which there are direct measurements of black-hole masses, and for which we can also determine their SFHs. There are 74 in total, probing total stellar masses from $M \approx 1 \times 10^{10} M_\odot$ to $M \approx 2 \times 10^{12} M_\odot$, where $M_\odot$ is the mass of the Sun. We removed from the final sample galaxies with strong nebular emission lines, in particular around the optical Hβ line, which affected the quality of the STECKMAP fit. Galaxies with prominent emission lines populate only the low-σ end of our sample ($\log\sigma \lesssim 2$). We normalized individual SFHs so that each galaxy has formed one mass unit at redshift $z \approx 0$.

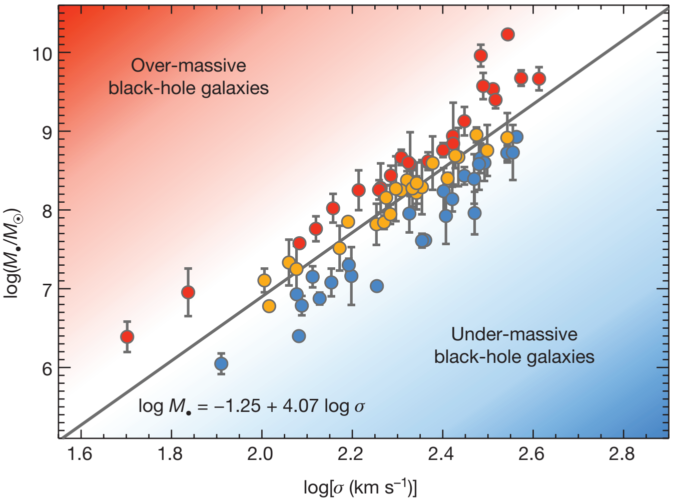


Figure 1 | Dispersion relation between black-hole mass and stellar velocity. The stellar velocity dispersion of galaxies (σ) tightly correlates with the mass of their supermassive black hole ($M_\bullet$). Data points correspond to the 74 HETMGS galaxies with measured black-hole masses and high-quality spectra. The solid line indicates the average black-hole mass for a given velocity dispersion. Galaxies more than +0.2 dex above this best-fitting $M_\bullet$–σ relation have black holes more massive than expected for their velocity dispersion, and therefore are called over-massive black-hole galaxies (red). Conversely, galaxies hosting less-massive black holes than the average population (by −0.2 dex or beyond) are called under-massive black-hole galaxies (blue). Galaxies with standard black-hole masses are shown in orange. Error bars are 1σ uncertainties.

[1]University of California Observatories, 1156 High Street, Santa Cruz, California 95064, USA. [2]Max-Planck Institut für Astronomie, Konigstuhl 17, D-69117 Heidelberg, Germany. [3]Department of Physics and Astronomy, San José State University, One Washington Square, San Jose, California 95192, USA. [4]Instituto de Astrofísica de Canarias, E-38205 La Laguna, Tenerife, Spain. [5]Departamento de Astrofísica, Universidad de La Laguna, E-38200 La Laguna, Tenerife, Spain. [6]European Southern Observatory, Karl-Schwarzschild-Strasse 2, 85748 Garching bei München, Germany.

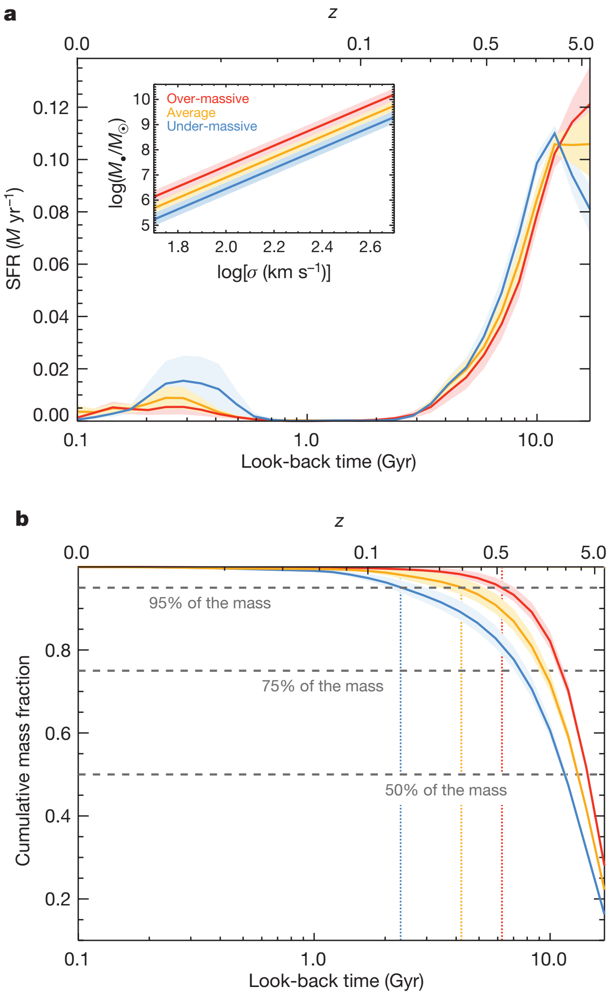


Figure 2 | Evolution of star formation over cosmic time. Red, orange and blue solid lines correspond to over-massive, standard and under-massive black-hole galaxies, respectively. Shaded regions indicate 1σ uncertainties in the mean values (but not model uncertainties). M, normalized mass. **a**, The evolution of the star formation rate (SFR) as a function of look-back time; **b**, the cumulative mass distribution for the three types of galaxy. Vertical dotted lines in **b** indicate when 95% of the final mass has been reached. The coupling between black-hole mass and star formation applies to all generations of stars, from those formed at $z \approx 5$ to the youngest generations. Our measurements of the SFH as a function of black-hole mass show how the latter drives baryonic cooling within massive haloes.

Our main result is summarized in Fig. 2, in which we show how star formation rates and cumulative stellar mass have evolved in over-massive (red), in standard (orange) and in under-massive (blue) black-hole galaxies. This evolution of star formation over cosmic time is strongly coupled to the mass of the central black hole. Galaxies with over-massive black holes experienced more intense star formation rates in the very early Universe (look-back times of 10 Gyr or more) than did galaxies with less-massive black holes. Star formation in over-massive black-hole galaxies was quenched earlier, with these galaxies reaching 95% of their final mass about 4 Gyr earlier, on average, than under-massive black-hole galaxies, as shown by the cumulative mass distributions. The amount of recent star formation, however, inversely correlates with the mass of the black holes: that is, young stellar populations are more prominent in under-massive black-hole galaxies. A Kolmogorov–Smirnov test of the distributions shown in Fig. 2 indicates that they are significantly different (P value of 0.026).

It is worth emphasizing that SFHs and black-hole masses are completely independent observables. Black-hole masses were calculated using a wide variety of methods but without detailed information on the stellar population properties[12]. If the differences presented in Fig. 2 were artefacts of the stellar population analysis, they could not be coupled to the mass of the black hole. This relative character of our approach minimizes the effect of systematic errors in the analysis. We have further checked that our choices for the STECKMAP free parameters do not affect our conclusions, and neither does the adopted $M_\bullet$–σ relation, nor the stellar population modelling (see Methods). Note also that there is no significant difference between the velocity dispersions of under-massive, standard and over-massive black-hole galaxies ($\log\sigma_{\rm under} = 2.32 \pm 0.04$, $\log\sigma_{\rm standard} = 2.30 \pm 0.03$, $\log\sigma_{\rm over} = 2.32 \pm 0.04$).

It could be argued that the process regulating black-hole growth also affected the efficiency of baryonic cooling within galaxies. In particular, objects formed in high-density environments could have grown more-massive black holes and formed their stellar populations differently owing to the amount and properties of the available gas. However, the lack of a morphological offset across the $M_\bullet$–σ relation[12,19] disfavours such a scenario (see also Methods). In addition, differences in formation timescales such as those shown in Fig. 2 do not depend on galaxy environment[20]. Thus, over-massive and under-massive black-hole galaxies have probably experienced similar formation paths. It is worth noting here the relative character of our analysis, that is, independent of the normalization of the $M_\bullet$–σ relation. Moreover, the robustness of our results with respect to additional parameters such as galaxy size or stellar density (see Methods) further indicates that galaxies with over-massive and under-massive black holes are different only in terms of their detailed stellar population properties and black-hole masses. Dynamically, morphologically and structurally, the two types of galaxy are indistinguishable.

The measurements shown in Fig. 2 probe the star formation processes within massive haloes since the early Universe. Interestingly, black-hole masses and star formation seem to be related as early as $z \approx 5$. This invalidates any scenario in which the observed scaling relations between black holes and host galaxies would emerge non-causally from the hierarchical evolution of a lambda cold-dark-matter (ΛCDM) Universe[21,22]. At the peak rate of star formation, baryon cooling was more efficient in galaxies with (present-day) more-massive black holes. The stellar mass formed around $z \approx 5$ in over-massive black-hole galaxies is about 1.3 times that formed in under-massive black-hole galaxies. Assuming the ratio of stellar mass to black-hole mass observed in the local Universe[12], these differences in the amount of stellar mass formed at $z \approx 5$ imply that more than 50% of the (vertical) scatter in the $M_\bullet$–σ relation results from this initial phase of galaxy formation and black-hole growth. We hypothesize that over-massive black-hole galaxies rapidly reached a black-hole mass capable of quenching star formation, which led to a shorter timescale for star formation. Thus, the baryon cooling efficiency at high redshift would play a major role in determining the present-day mass of supermassive black holes, feeding the primordial seeds of the black holes with gas, in agreement with quasar observations[23].

The importance of supermassive black holes in galaxy evolution arises from their potential role as quenching agents. We found that in those galaxies with less-massive central black holes, star formation lasted longer. This time delay, consistent with the observed differences in the abundance of alpha-process elements versus iron [α/Fe] of over-massive and under-massive black-hole galaxies[24], is naturally explained if quenching is driven by active galactic nucleus (AGN) feedback. Accretion onto higher-mass black holes leads to more energetic AGN feedback which would quench the star formation faster. This high-redshift picture has its $z \approx 0$ counterpart, as recent star formation is also expected to be regulated by AGN activity. If the rate of energy injection scales with the mass of the black hole[8,9], less-massive black holes, growing at low accretion rates in the nearby Universe, would be less efficient at keeping hot the gaseous corona, which will ultimately cool and form new stars[25]. In Fig. 2, the fraction of young stars anticorrelates with the relative mass of the black hole, further supporting an active role of black holes in regulating star formation within massive galaxies.

Investigating the connection between star formation and black-hole activity has been one of the biggest observational challenges since AGNs were proposed as the main source of feedback within massive galaxies. Whereas star formation takes place over long periods of time, the rapid and nonlinear response of black holes to gas accretion[26] complicates a clean empirical comparison between AGN luminosity and star formation rate. AGNs typically populate star-forming galaxies, but their luminosities may not correlate with observed rates of star formation[27–30]. Here, we have made use of the relation between black-hole mass and SFHs to show that the evolution of star formation in massive galaxies over cosmic time is driven by black-hole activity. Our results indicate that there may be a causal origin for the observed scaling relations between galaxy properties and black-hole mass, offering observational support for AGN-based quenching mechanisms.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 May; accepted 28 October 2017.
Published online 1 January 2018.

1. Magorrian, J. *et al.* The demography of massive dark objects in galaxy centers. *Astron. J.* **115,** 2285–2305 (1998).
2. Gebhardt, K. *et al.* A relationship between nuclear black hole mass and galaxy velocity dispersion. *Astrophys. J.* **539,** L13–L16 (2000).
3. Silk, J. & Mamon, G. A. The current status of galaxy formation. *Res. Astron. Astrophys.* **12,** 917–946 (2012).
4. Vogelsberger, M. *et al.* Introducing the Illustris Project: simulating the coevolution of dark and visible matter in the Universe. *Mon. Not. R. Astron. Soc.* **444,** 1518–1547 (2014).
5. Schaye, J. *et al.* The EAGLE project: simulating the evolution and assembly of galaxies and their environments. *Mon. Not. R. Astron. Soc*. **446,** 521–554 (2015).
6. Choi, E. *et al.* Physics of galactic metals: evolutionary effects due to production, distribution, feedback & interaction with black holes. *Astrophys. J.* **844,** 31 (2016).
7. Ferrarese, L. & Merritt, D. A fundamental relation between supermassive black holes and their host galaxies. *Astrophys. J.* **539,** L9–L12 (2000).
8. Crain, R. A. *et al.* The EAGLE simulations of galaxy formation: calibration of subgrid physics and model variations. *Mon. Not. R. Astron. Soc*. **450,** 1937–1961 (2015).
9. Sijacki, D. *et al.* The Illustris simulation: the evolving population of black holes across cosmic time. *Mon. Not. R. Astron. Soc.* **452,** 575–596 (2015).
10. van den Bosch, R. C. E., Gebhardt, K., Gültekin, K., Yldrm, A. & Walsh, J. L. Hunting for supermassive black holes in nearby galaxies with the Hobby–Eberly Telescope. *Astrophys. J. Suppl. Ser.* **218,** 10 (2015).
11. Rodriguez-Gomez, V. *et al.* The stellar mass assembly of galaxies in the Illustris simulation: growth by mergers and the spatial distribution of accreted stars. *Mon. Not. R. Astron. Soc.* **458,** 2371–2390 (2016).
12. van den Bosch, R. C. E. Unification of the fundamental plane and super massive black hole masses. *Astrophys. J.* **831,** 134 (2016).
13. Ocvirk, P., Pichon, C., Lançon, A. & Thiébaut, E. STECKMAP: STEllar Content and Kinematics from high resolution galactic spectra via Maximum A Posteriori. *Mon. Not. R. Astron. Soc*. **365,** 74–84 (2006).
14. Vazdekis, A. *et al.* Evolutionary stellar population synthesis with MILES—I. The base models and a new line index system. *Mon. Not. R. Astron. Soc.* **404,** 1639–1671 (2010).
15. Koleva, M., Prugniel, P., Ocvirk, P., Le Borgne, D. & Soubiran, C. Spectroscopic ages and metallicities of stellar populations: validation of full spectrum fitting. *Mon. Not. R. Astron. Soc.* **385,** 1998–2010 (2008).
16. Sánchez-Blázquez, P., Ocvirk, P., Gibson, B. K., Pérez, I. & Peletier, R. F. Star formation history of barred disc galaxies. *Mon. Not. R. Astron. Soc.* **415,** 709–731 (2011).
17. Leitner, S. N. On the last 10 billion years of stellar mass growth in star-forming galaxies. *Astrophys. J.* **745,** 149 (2012).
18. Ruiz-Lara, T. *et al.* Recovering star formation histories: Integrated-light analyses vs. stellar colour-magnitude diagrams. *Astron. Astrophys.* **583,** A60 (2015).
19. Beifiori, A., Courteau, S., Corsini, E. M. & Zhu, Y. On the correlations between galaxy properties and supermassive black hole mass. *Mon. Not. R. Astron. Soc.* **419,** 2497–2528 (2012).
20. Thomas, D., Maraston, C., Bender, R. & Mendes de Oliveira, C. The epochs of early-type galaxy formation as a function of environment. *Astrophys. J.* **621,** 673–694 (2005).
21. Peng, C. Y. How mergers may affect the mass scaling relation between gravitationally bound systems. *Astrophys. J.* **671,** 1098–1107 (2007).
22. Jahnke, K. & Macciò, A. V. The non-causal origin of the black-hole-galaxy scaling relations. *Astrophys. J.* **734,** 92 (2011).
23. Fan, X. *et al.* A survey of $z>5.8$ quasars in the Sloan Digital Sky Survey. I. Discovery of three new quasars and the spatial density of luminous quasars at $z=6$. *Astron. J.* **122,** 2833–2849 (2001).
24. Martin-Navarro, I., Brodie, J. P., van den Bosch, R. C. E., Romanowsky, A. J. & Forbes, D. A. Stellar populations across the black hole mass-velocity dispersion relation. *Astrophys. J.* **832,** L11 (2016).
25. Terrazas, B. A. *et al.* Quiescence correlates strongly with directly measured black hole mass in central galaxies. *Astrophys. J.* **830,** L12 (2016).
26. Bower, R. G. *et al.* The dark nemesis of galaxy formation: why hot haloes trigger black hole growth and bring star formation to an end. *Mon. Not. R. Astron. Soc.* **465,** 32–44 (2017).
27. Kauffmann, G. *et al.* The host galaxies of active galactic nuclei. *Mon. Not. R. Astron. Soc*. **346,** 1055–1077 (2003).
28. Mullaney, J. R. *et al.* GOODS-Herschel: the far-infrared view of star formation in active galactic nucleus host galaxies since $z\approx3$. *Mon. Not. R. Astron. Soc.* **419,** 95–115 (2012).
29. Stanley, F. *et al.* A remarkably flat relationship between the average star formation rate and AGN luminosity for distant X-ray AGN. *Mon. Not. R. Astron. Soc.* **453,** 591–604 (2015).
30. Ruschel-Dutra, D., Rodríguez Espinosa, J. M., González Martín, O., Pastoriza, M. & Riffel, R. Star formation in AGNs at the hundred parsec scale using MIR high-resolution images. *Mon. Not. R. Astron. Soc.* **466,** 3353–3363 (2017).

Acknowledgements We acknowledge support from the National Science Foundation (NSF) grants AST-1616598 and AST-1616710, from a Marie Curie Global Fellowship, from SFB 881 'The Milky Way System' (subprojects A7 and A8) funded by the German Research Foundation, and from grants AYA2016-77237-C3-1-P and AYA2014-56795-P from the Spanish Ministry of Economy and Competitiveness (MINECO). G.v.d.V. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement no. 724857 (Consolidator Grant ArcheoDyn). A.J.R. was supported as a Research Corporation for Science Advancement Cottrell Scholar. I.M.-N. thanks the TRACES group and M. Mezcua for their comments, R. van den Bosch for making the HETMGS data publicly available and A. Boecker for her help.

Author Contributions I.M.-N. derived the star formation histories along with T.R.-L. and wrote the text. J.P.B., A.J.R., T.R.-L. and G.v.d.V. contributed to the interpretation and analysis of the results.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to I.M.-N. (imartinn@ucsc.edu).

Reviewer Information *Nature* thanks D. Rosario, J. Trump and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data quality and spectral fitting. Representative spectra of a low-σ and a high-σ galaxy are shown in Extended Data Fig. 1, with the best-fitting STECKMAP model overplotted. Because of the high signal-to-noise ratio of the data (typically above about 100 Å^{-1}), the residuals are typically below 1% (approximately 0.06 and 0.08 for these low-σ and high-σ objects, respectively).

Velocity dispersion–metallicity degeneracy. There is a well-known degeneracy between σ and galaxy metallicity[15], which could potentially affect our measurements of the SFH. It has been shown that fitting kinematics and stellar population properties independently minimizes the effect of this degeneracy[16]. Thus, we first determined the kinematics (systemic velocity $V_{\rm sys}$ and σ) using the penalized pixel-fitting method (pPXF)[31], which was also used to remove the nebular emission from our spectra. The temporal combination of models of single stellar populations. convolved to the resolution of the galaxy measured with pPXF, was used to calculate the SFHs.

As a further test to assess the dependence of our results on the adopted velocity dispersion, we repeated the analysis while allowing STECKMAP to measure the kinematics at the same time as the SFHs, although this approach has been proved to be less accurate[16]. In Extended Data Fig. 2, we show the cumulative mass distributions of under-massive and over-massive black-hole galaxies measured in this way. The observed differences in the SFHs across the relation between black-hole mass and σ are not due to degeneracies between stellar populations and kinematical properties.

Robustness of the results. *Regularization parameters.* To assess the robustness of our results, we varied the two main free parameters in our analysis. On the one hand, STECKMAP allows for a regularization in both the SFH (μ_x) and the age–metallicity relation (μ_Z) of the different stellar population models. Effectively, this regularization behaves as a Gaussian prior[32]. Figure 2 was calculated using $\mu_x = \mu_Z = 10$. Although the choice of these parameters mainly depends on the quality and characteristics of the observed spectra, we repeated the analysis but varying each regularization parameter by two orders of magnitude. In Extended Data Fig. 3, we demonstrate that our choice of the regularization parameters does not affect the main conclusions of this work but provides the most stable solutions within the range of μ_x and μ_Z values explored.

Sample selection. As described in the main text, the final sample consists of every galaxy in the HETMGS survey with good enough spectra to perform our stellar population analysis. In practice, this means rejecting spectra with very low signal-to-noise ratio (less than about 10) and strong emission lines. Possible biases related to these selection criteria are discussed below. No additional constraints were applied. Our best-fitting relation is based on our own determinations of the velocity dispersion of individual galaxies, homogeneously measured at half $R_{\rm e}$. Galaxies with black-hole masses departing from our best-fitting $M_\bullet$–σ by more than +0.2 or −0.2 dex were classified as over-massive or under-massive black-hole galaxies, respectively. With this criterion, the number of over-massive, standard and under-massive black-hole galaxies is similar, at 25, 24 and 25 objects, respectively. As reported in previous studies of large sample of black-hole masses[12,19], there is no morphological dependence of our best-fitting $M_\bullet$–σ relation. In Extended Data Fig. 4, we show the distributions for the concentration parameter $C_{28} \equiv 5\log(R_{80}/R_{20})$ for under-massive and over-massive black-hole galaxies. Galactocentric distances R_{80} and R_{20} encompass 80% and 20% of the total light of the galaxy, respectively, and were measured using elliptical isophotes[12]. The parameter C_{28} is a proxy for the light concentration of galaxies and therefore of their morphology. Higher C_{28} corresponds to ellipsoids, whereas lower values are associated with galaxies that are more disk-like[33,34]. As expected, C_{28} behaves similarly across the $M_\bullet$–σ relation, with median values of 5.5 and 5.6 for over-massive and under-massive black-hole galaxies, respectively. A two-sided Kolmogorov–Smirnov test indicates that the differences between the two C_{28} distributions are insignificant ($P = 0.82$). Thus, over-massive black-hole galaxies are morphologically indistinguishable from under-massive black-hole objects.

We also investigated whether the assumed best-fitting $M_\bullet$–σ relation could lead to spurious results. As an extreme test, we recalculated the mean SFHs but, instead of using our best-fitting solution, we assumed the one calculated for a much larger sample of black-hole masses[12]. This sample includes objects that were not observed by HETMGS or whose spectra were rejected in our analysis because of the poor quality. Specifically, this alternative $M_\bullet$–σ relation is given by $\log M_\bullet = -4.00 + 5.35\log\sigma$. Note that the use of this equation is not consistent with our sample. Velocity dispersions were calculated differently and over different radial apertures ($0.5R_{\rm e}$ versus $1R_{\rm e}$). Thus, adopting this equation to distinguish between over-massive and under-massive black-hole galaxies could potentially affect our conclusions. Despite this, Extended Data Fig. 3 shows that the dependence of the SFH on the mass of the central black hole stands, regardless of the implicit $M_\bullet$–σ relation.

It is worth noting that the zero point and the slope of the $M_\bullet$–σ relation depend on the effective velocity dispersion of individual galaxies and, to a lesser extent, on the assumed mass of the black hole. Whereas the latter is unambiguously defined, the stellar velocity dispersion varies considerably within galaxies. However, the relative distinction between over-massive and under-massive black-hole galaxies is relatively insensitive to the characteristics of different samples. The majority of objects (about 85%) classified as over-massive or under-massive using our best-fitting relation are also over-massive or under-massive according to other widely adopted $M_\bullet$–σ relations[12,35]. Thus, our classification and therefore our conclusions do not depend on absolute σ values.

Finally, we have also considered the possibility that differences in the internal kinematics or morphology of galaxies could bias our conclusions. In particular, under-massive black-hole galaxies may be offset from the average $M_\bullet$–σ relation owing to a higher prevalence of disk-like, rotationally supported structures, which in general tend to show younger populations[36]. Observationally, there is no evidence of a morphological or mass-concentration dependency of the $M_\bullet$–σ relation[19], although it has been claimed that pseudo-bulges may follow a different scaling relation[19,37,38], which becomes relevant for velocity dispersions $\sigma < 200\,{\rm km\,s^{-1}}$. In Extended Data Fig. 5, we show the star formation rate as a function of look-back time only for galaxies above this velocity dispersion threshold, where the bulk of the population is dispersion-supported. It is clear that the observed differences in the star formation processes between over-massive and under-massive black-hole galaxies are not due to a morphological or kinematical effect.

Nebular emission and young stellar populations. Although the age sensitivity of the spectra is widely spread over the whole wavelength range[32], our strongest age-sensitive feature is the Hβ line. Thus, we decided to be conservative and remove from the analysis objects with strong emission lines (amplitude-to-noise ratio > 4), which effectively leads to the old stellar populations shown in Fig. 2. If we include galaxies with stronger emission lines, STECKMAP recovers the expected contribution of younger stars, without softening the differences between over-massive and under-massive black-hole galaxies, as shown in Extended Data Fig. 6.

Stellar population synthesis models. The main source of systematic uncertainties in any stellar population analysis is the choice of a given set of models. We used the MILES stellar population models as a reference, as they provide fully empirical spectroscopic predictions over the explored wavelength range and are best suited for intermediate-to-old stellar populations[14]. However, given our relatively narrow coverage of wavelengths, we also addressed the impact of the choice of stellar population model on our results. We repeated the analysis of our sample of galaxies but considering three additional sets of models, namely the PÉGASE-HR model[39], the Bruzual and Charlot 2003 (BC03) model[40] and the GRANADA/MILES model[41,42]. As shown in Extended Data Fig. 7, the differences between over-massive and under-massive black-hole galaxies are clear in all three models. The agreement between different models is remarkable given the strong underlying differences among them. As an additional test of the robustness of our analysis, we also compared our observations to a new set of the MILES models which make use of BaSTi isochrones[43]. This final comparison is also included in Extended Data Fig. 7, further reinforcing the conclusion that our results are not driven by systematics in the stellar population analysis.

Galaxy densities and sample biases. The statistical properties of galaxies with measured black-hole masses do not follow those of the overall population of galaxies. In particular, objects with known black-hole masses tend to be denser than the average[10,44], partially because our ability to measure black-hole masses depends on how well we can resolve their spheres of influence[45]. To test whether density is a confounding variable in our analysis, we performed a bilinear fitting between $M_\bullet$–σ and stellar density ($M_\star/R_{\rm e}^3$). We used publicly available measurements of sizes ($R_{\rm e}$) and K-band luminosities[12] to obtain a best-fitting relation given by

$$\log M_\bullet = -1.5 + 4.42\log\sigma - 0.05\log(M_\star R_{\rm e}^{-3}) \qquad (1)$$

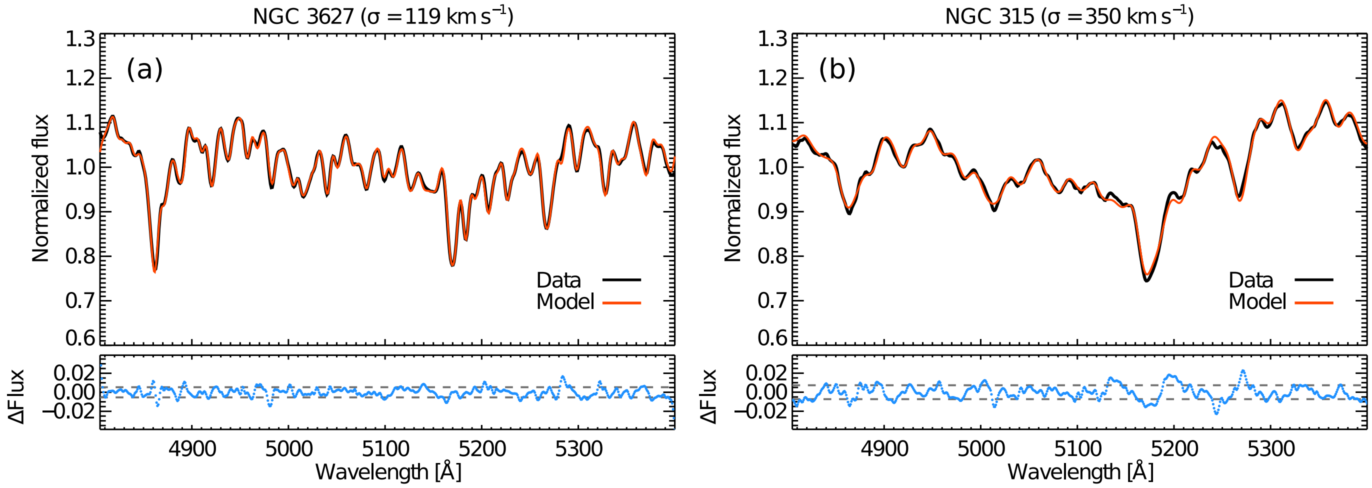
where $M_\star$ is the stellar mass based on the K-band luminosity and $R_{\rm e}$ the effective radius. This best-fitting relation, consistent with previous studies[46] (Extended Data Fig. 8), allowed us to investigate the effect of the black hole at fixed stellar velocity dispersion and density. This relation is still mostly driven by σ and only weakly depends on stellar density. The cumulative mass distribution of over-massive and under-massive black-hole galaxies, according to the equation above, is shown in Extended Data Fig. 8. We found that, when stellar density is also taken into account, the decoupling between over-massive and under-massive black-hole galaxies remains unaltered. Thus, the differences observed in the SFHs are not due to different galaxy densities. Selection biases in the current sample of black-hole masses would, if anything[47], change only the normalization of the $M_\bullet$–σ relation[45]. Our approach, by construction, is insensitive to this effect.

Additionally, we explored possible bias introduced by our rejection criteria—that is, those galaxies that we did not include in the analysis either because of low signal-to-noise ratio or because of strong emission features. In Extended Data Fig. 9, we show the R_e–σ distribution for HETMGS galaxies, for which the sizes were taken from the 2MASS Extended Source catalogue[48], averaged following the method adopted by the ATLAS^{3D} team[49]. We used the stellar velocity dispersions listed in the HETMGS presentation paper[10]. The typical sizes of over-massive and under-massive black-hole galaxies ($R_e^{OM} = 2.69 \pm 0.34$ kpc and $R_e^{UM} = 2.49 \pm 0.37$ kpc, respectively) are consistent with the average population of galaxies with known black-hole masses (R_e^{BH}=2.74 ± 0.22 kpc). The same applies to the K-band luminosities, with a typical value of $\log(L_K^{OM}/L_\odot) = 11.13 \pm 0.08$, $\log(L_K^{UM}/L_\odot) = 11.05 \pm 0.07$ and $\log(L_K^{BH}/L_\odot) = 11.08 \pm 0.05$, for over-massive black-hole galaxies, under-massive black-hole galaxies and the complete sample of black-hole galaxies, respectively. We therefore conclude that our final sample is not significantly biased relative to the average population of galaxies with measured black-hole masses.

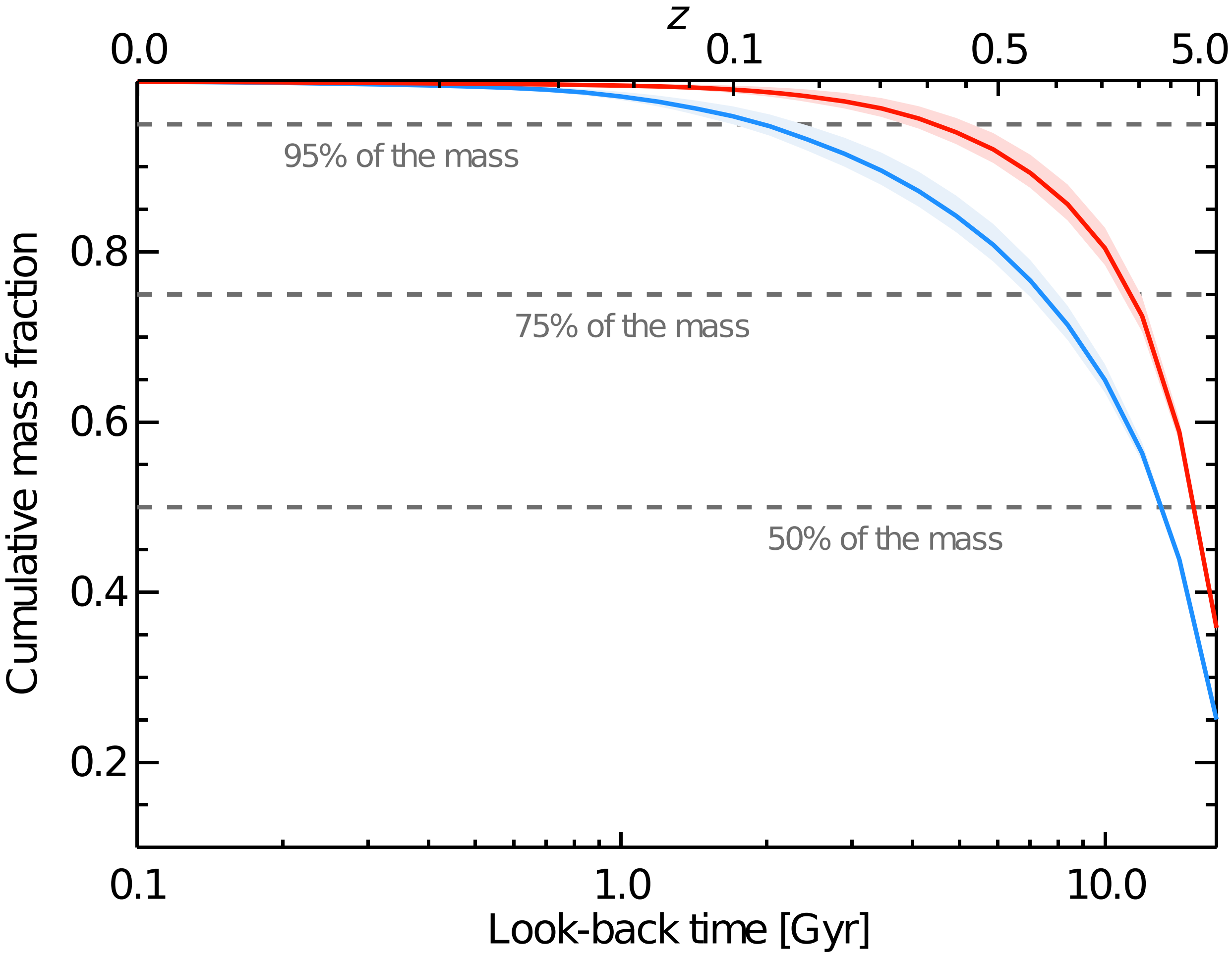
Code availability. The STECKMAP code used to derive the SFHs is publicly available at http://astro.u-strasbg.fr/~ocvirk/indexsteckmap.html.

Data availability. All data analysed during this study are available at the HETMGS website http://www.mpia.de/~bosch/hetmgs/.

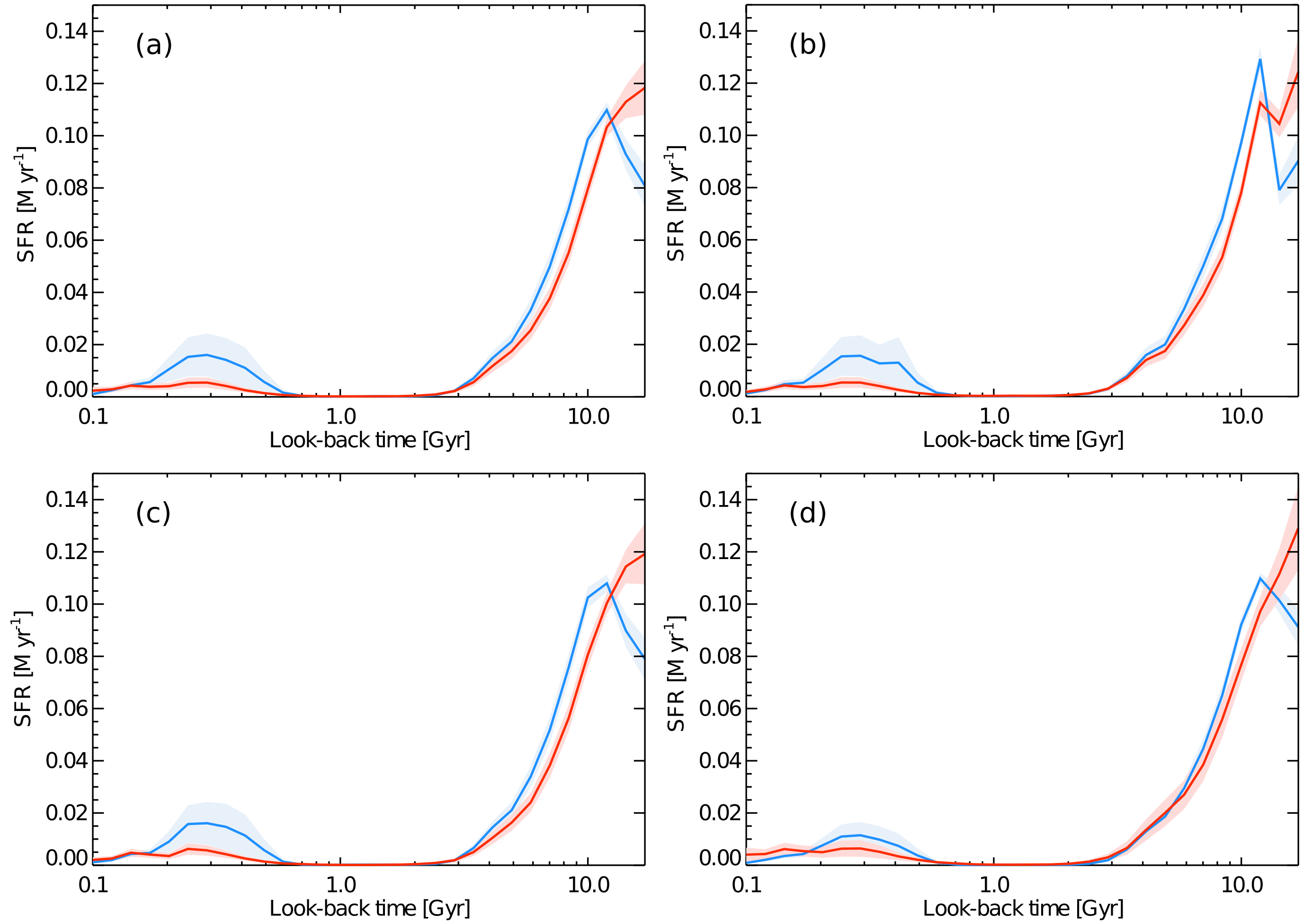
31. Cappellari, M. & Emsellem, E. Parametric recovery of line-of-sight velocity distributions from absorption-line spectra of galaxies via penalized likelihood. *Publ. Astron. Soc. Pacif.* **116,** 138–147 (2004).
32. Ocvirk, P., Pichon, C., Lançon, A. & Thiébaut, E. STECMAP: STEllar Content from high-resolution galactic spectra via Maximum A Posteriori. *Mon. Not. R. Astron. Soc.* **365,** 46–73 (2006).
33. Shimasaku, K. *et al.* Statistical properties of bright galaxies in the Sloan Digital Sky Survey Photometric System. *Astron. J.* **122,** 1238–1250 (2001).
34. Courteau, S., McDonald, M., Widrow, L. M. & Holtzman, J. The bulge–halo connection in galaxies: a physical interpretation of the V_c–σ_0 relation. *Astrophys. J.* **655,** L21–L24 (2007).
35. Kormendy, J. & Ho, L. C. Coevolution (or not) of supermassive black holes and host galaxies. *Annu. Rev. Astron. Astrophys.* **51,** 511–653 (2013).
36. Kuntschner, H. *et al.* The SAURON project: XVII. Stellar population analysis of the absorption line strength maps of 48 early-type galaxies. *Mon. Not. R. Astron. Soc.* **408,** 97–132 (2010).
37. Hu, J. The black hole mass–stellar velocity dispersion correlation: bulges versus pseudo-bulges. *Mon. Not. R. Astron. Soc.* **386,** 2242–2252 (2008).
38. Kormendy, J., Bender, R. & Cornell, M. E. Supermassive black holes do not correlate with galaxy disks or pseudobulges. *Nature* **469,** 374–376 (2011).
39. Le Borgne, D. *et al.* Evolutionary synthesis of galaxies at high spectral resolution with the code PEGASE-HR. Metallicity and age tracers. *Astron. Astrophys.* **425,** 881–897 (2004).
40. Bruzual, G. & Charlot, S. Stellar population synthesis at the resolution of 2003. *Mon. Not. R. Astron. Soc.* **344,** 1000–1028 (2003).
41. González Delgado, R. M., Cerviño, M., Martins, L. P., Leitherer, C. & Hauschildt, P. H. Evolutionary stellar population synthesis at high spectral resolution: optical wavelengths. *Mon. Not. R. Astron. Soc.* **357,** 945–960 (2005).
42. González Delgado, R. M. & Cid Fernandes, R. Testing spectral models for stellar populations with star clusters: II. Results. *Mon. Not. R. Astron. Soc.* **403,** 797–816 (2010).
43. Pietrinferni, A., Cassisi, S., Salaris, M. & Castelli, F. A large stellar evolution database for population synthesis studies. I. Scaled solar models and isochrones. *Astrophys. J.* **612,** 168–190 (2004).
44. Bernardi, M., Sheth, R. K., Tundo, E. & Hyde, J. B. Selection bias in the $M_\bullet$–σ and $M_\bullet$–L correlations and its consequences. *Astrophys. J.* **660,** 267–275 (2007).
45. Shankar, F. *et al.* Selection bias in dynamically measured supermassive black hole samples: its consequences and the quest for the most fundamental relation. *Mon. Not. R. Astron. Soc.* **460,** 3119–3142 (2016).
46. Saglia, R. P. *et al.* The SINFONI Black Hole Survey: the black hole fundamental plane revisited and the paths of (co)evolution of supermassive black holes and bulges. *Astrophys. J.* **818,** 47 (2016).
47. Gültekin, K., Tremaine, S., Loeb, A. & Richstone, D. O. Observational selection effects and the M–σ relation. *Astrophys. J.* **738,** 17 (2011).
48. Jarrett, T. H. *et al.* 2MASS Extended Source Catalog: overview and algorithms. *Astron. J.* **119,** 2498–2531 (2000).
49. Cappellari, M. *et al.* The ATLAS^{3D} project: I. A volume-limited sample of 260 nearby early-type galaxies: science goals and selection criteria. *Mon. Not. R. Astron. Soc.* **413,** 813–836 (2011).



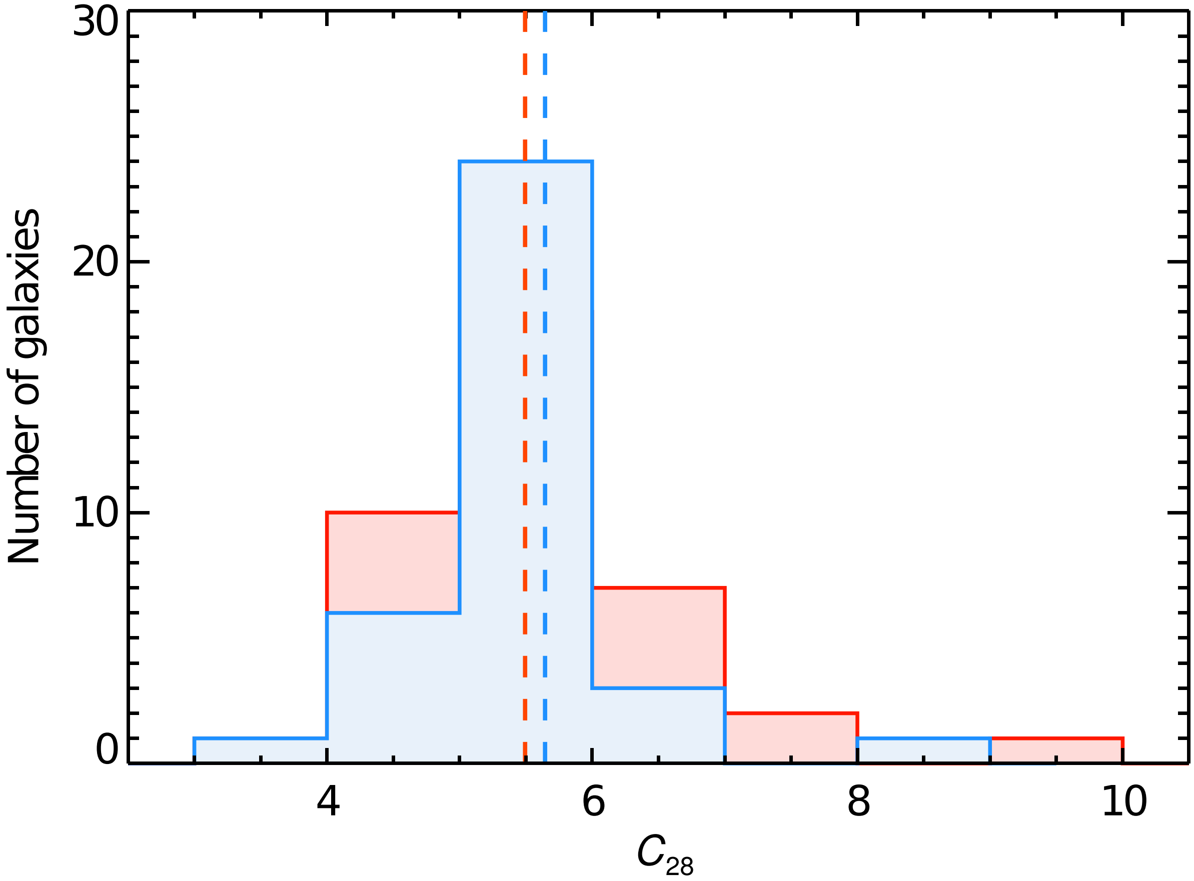
Extended Data Figure 1 | Data and best-fitting stellar populations model. **a**, The spectrum of the low-σ galaxy NGC 3627; **b**, the spectrum of the higher-σ galaxy NGC 315. Along with the HETMGS spectra (black line), we also show the best-fitting STECKMAP model (red line). In the bottom panel, we show the residuals (blue dots), which are in both cases below 2%. The standard deviation is shown as dashed horizontal lines.



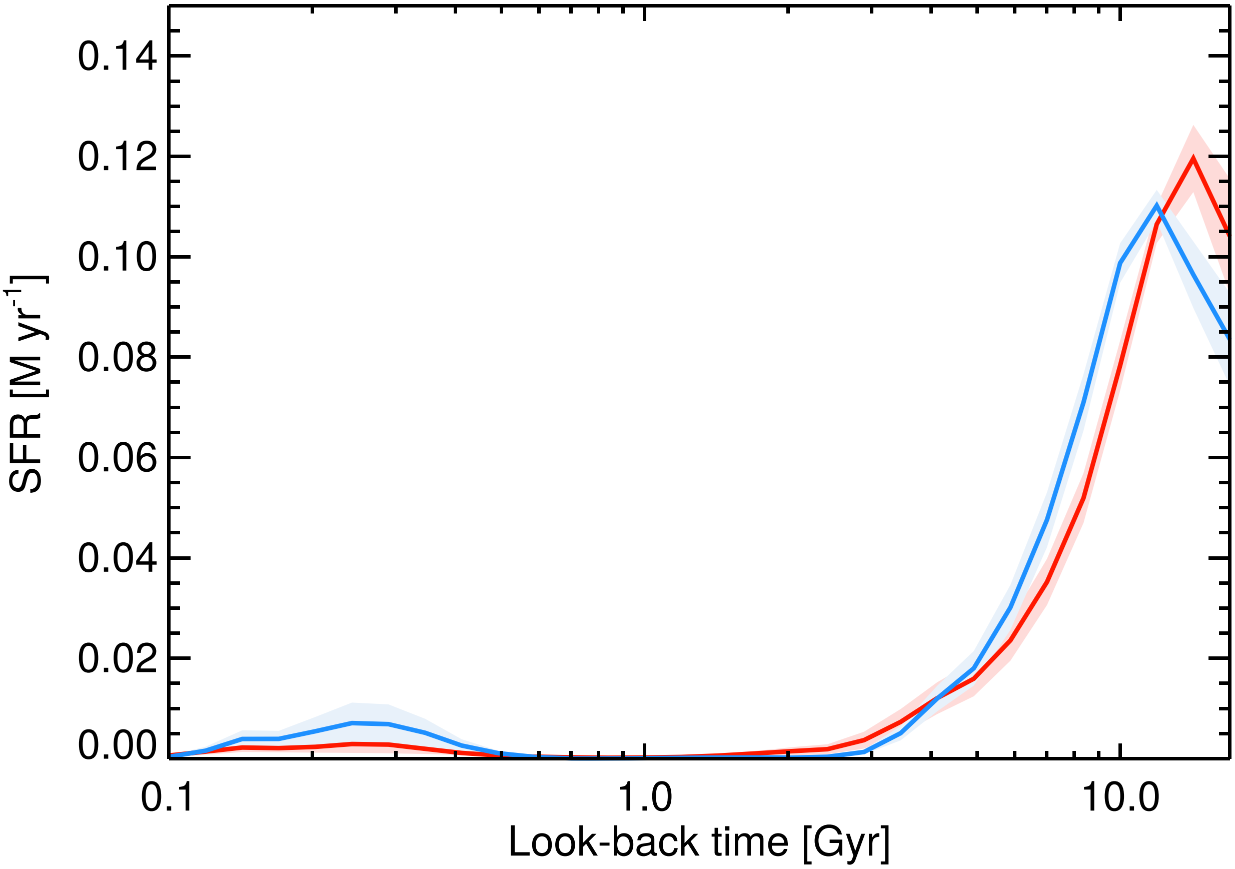
Extended Data Figure 2 | Cumulative mass distributions without fixing the kinematics. To assess the degeneracy between stellar population properties and stellar velocity dispersion, we repeated the analysis while allowing STECKMAP to fit the kinematics simultaneously with the SFHs. In red and blue, the cumulative mass fractions of under-massive (blue) and over-massive (red) black-hole galaxies are shown as a function of look-back time. Although leaving the kinematics as free parameters leads to less accurate SFHs[16], the differences in the star formation of under-massive and over-massive black-hole galaxies remain clear.



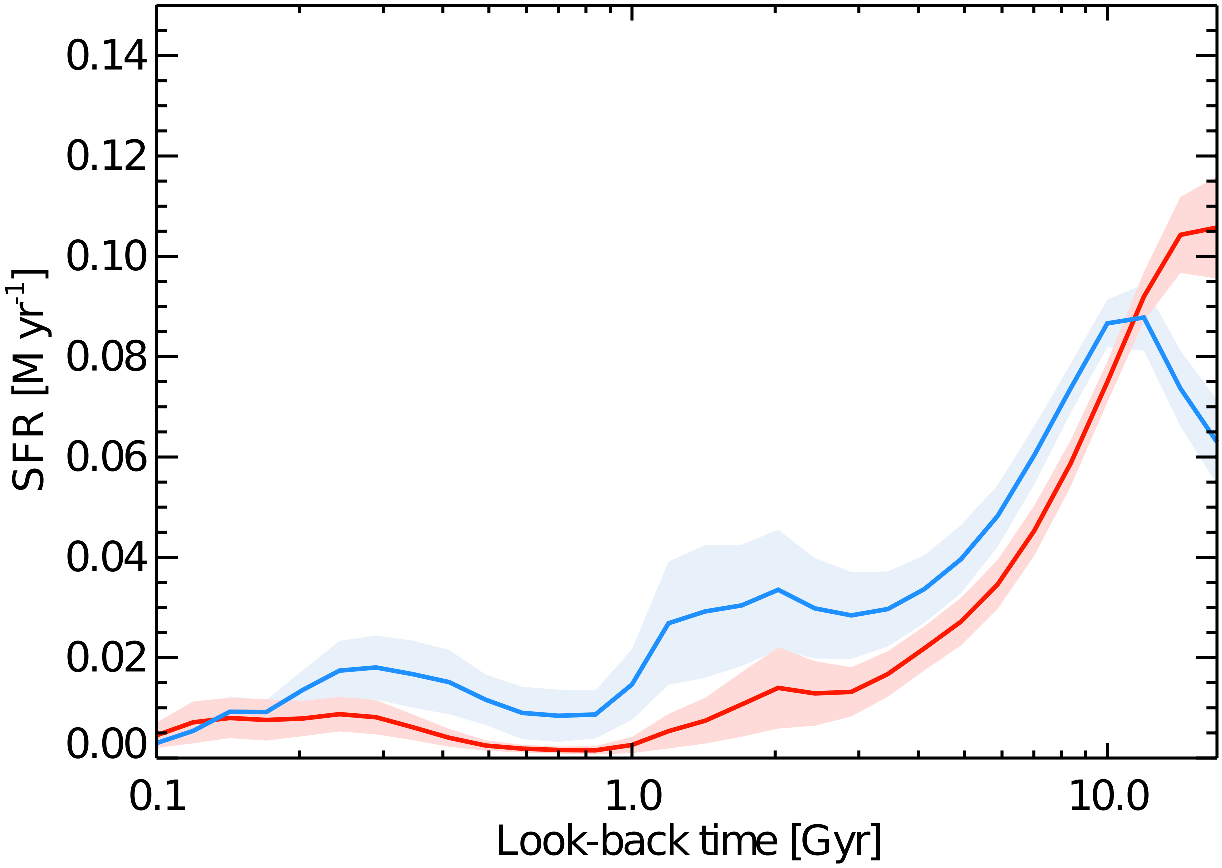
Extended Data Figure 3 | Robustness of the recovered star formation rates. Different panels correspond to the different tests performed in order to explore the reliability of our results. As in Fig. 2, red and blue lines indicate the star formation rate (SFR) as a function of look-back time for over-massive and under-massive black-hole galaxies, and the shaded areas mark the 1σ uncertainties. **a**, Our preferred model, as a reference. A two-sided Kolmogorov–Smirnov comparison between over-massive and under-massive galaxies indicates that the two distributions are significantly different ($P = 0.026$). **b**, Here, we left the regularization of the SFH almost free, by setting $\mu_x = 0.1$. **c**, We varied the regularization of the age–metallicity relation in the same way, changing μ_Z from 10 to 0.1. **d**, Finally, we adopted a different (but inconsistent) $M_\bullet$–σ relation[12] to separate our sample. All these tests demonstrate that our conclusions are insensitive to systematics in the analysis.



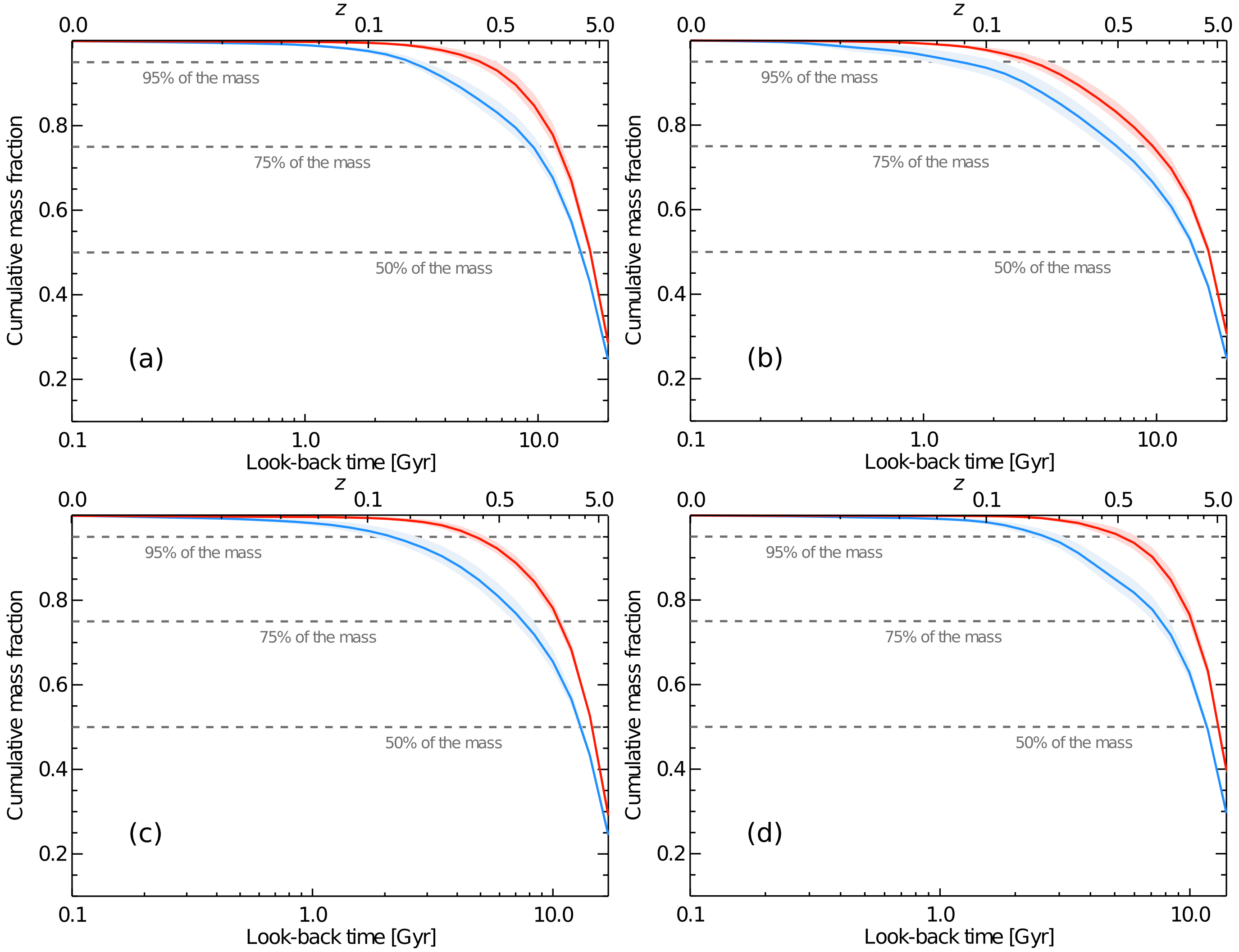
Extended Data Figure 4 | Distributions of concentration parameters. The distribution of the C_{28} concentration parameter is shown in red and blue for over-massive and under-massive black-hole galaxies, respectively. Higher (lower) concentration indices are associated with earlier (later) morphological types. Vertical dashed lines mark the median of the distributions. No significant differences are found between the samples, indicating that the over-massive and under-massive black-hole galaxies are morphologically indistinguishable. This suggests that both types of galaxy share the same formation processes but differ in the present-day mass of their central black holes.



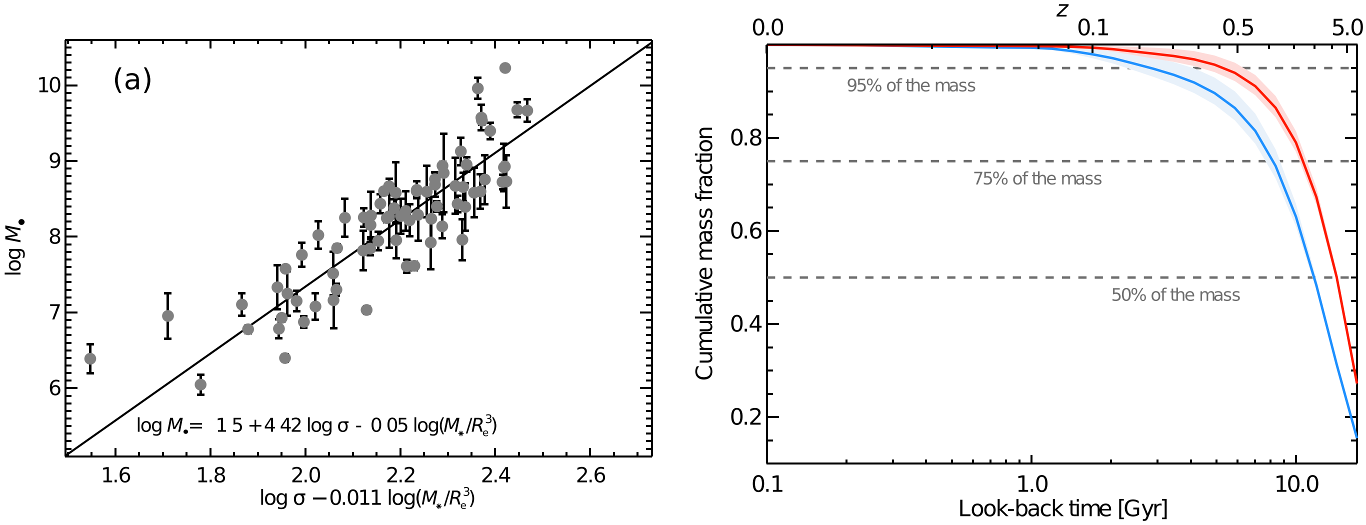
Extended Data Figure 5 | Star formation rate for high-mass galaxies. Differences in the star formation rate as a function of look-back time for over-massive (red) and under-massive (blue) black-hole galaxies, but only including objects with velocity dispersions $\sigma > 200$ km s^{-1} (log$\sigma = 2.32$). The massive end of the $M_\bullet$–σ relation is dominated by elliptical, pressure-supported systems[12,19] but shows the same distinction between over-massive and under-massive black-hole galaxies.



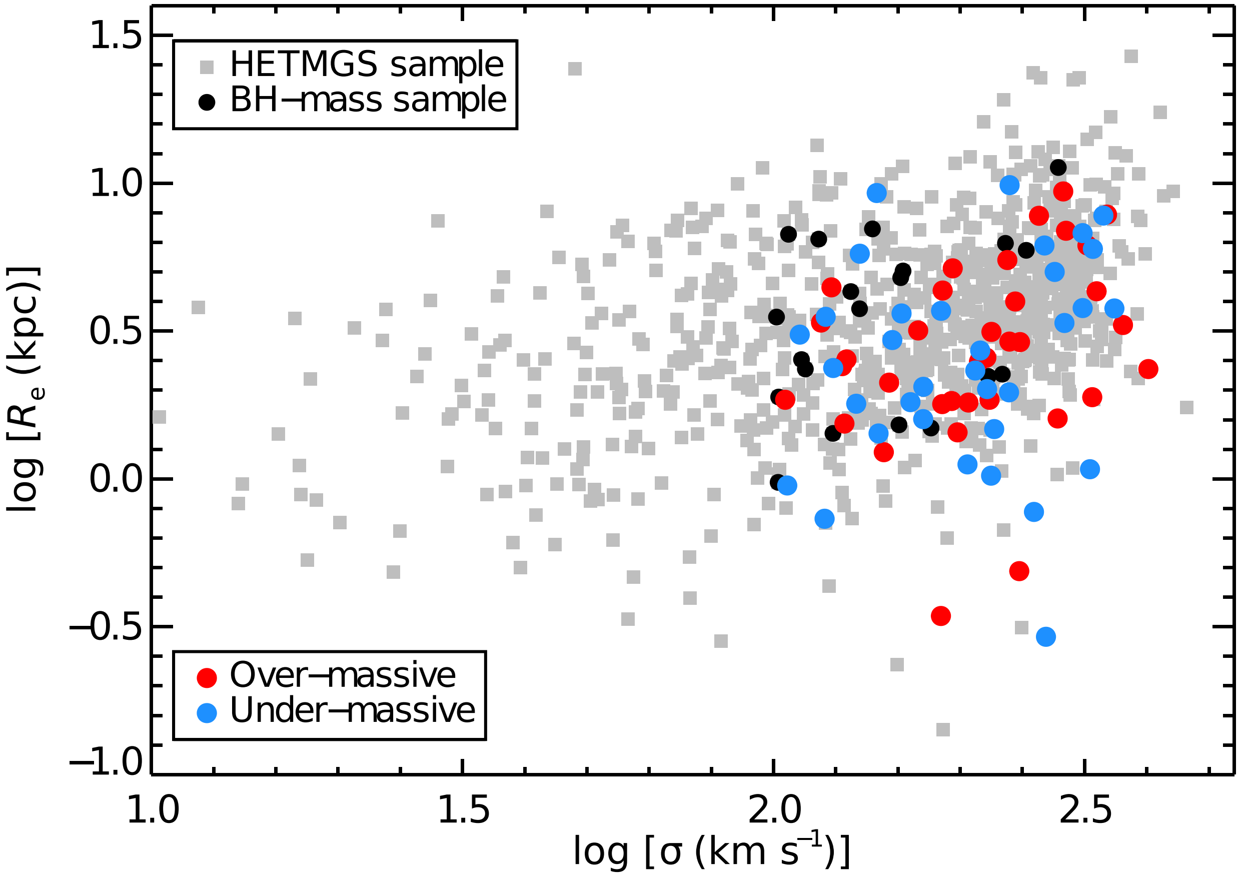
Extended Data Figure 6 | Nebular emission and star formation rate. If we include galaxies with strong emission lines in the analysis, the quality of the fits worsens but over-massive (red) and under-massive (blue) black-hole galaxies still show decoupled SFHs.



Extended Data Figure 7 | Dependence on the stellar population models. As in Fig. 2, each panel shows the cumulative mass fraction for over-massive (red) and under-massive (blue) black-hole galaxies. In each panel, the SFHs were calculated using a different set of stellar population synthesis models. **a**, Result based on the PÉGASE-HR[39] models; **b**, result based BC03 model[40]; **c**, result using the GRANADA/MILES models[41,42]; **d**, finally, result using the MILES models with BaSTi isochrones[43]. Despite the quantitative differences, the different behaviour of under-massive and over-massive black-hole galaxies is clear in all panels, and thus it is not an artefact of a particular set of models. Note that different models are fed with different stellar libraries, interpolated in different ways to populate different isochrones. The separation between over-massive and under-massive black-hole galaxies is thus independent of the different ingredients in stellar population modelling.



Extended Data Figure 8 | Stellar density as a confounding variable. **a**, $M_\bullet$ as a function of the best-fitting combination of σ and stellar density, following equation (1). Projecting over this plane allows us to separate galaxies depending on the mass of their black holes, but at fixed stellar velocity dispersion and stellar density ($M_\star/R_e^3$). **b**, The cumulative mass distributions of over-massive and under-massive black-hole galaxies according to this new definition. This test demonstrates that the observed differences in the SFHs are not due to a possible variation of the stellar density across the $M_\bullet$–σ relation, further supporting the role of the black hole in regulating star formation within massive galaxies.



Extended Data Figure 9 | Selection function. Filled grey squares show the size–*σ* relation for the HETMGS sample, and filled circles correspond to those galaxies with measured black-hole masses (blue and red indicate under-massive and over-massive, respectively), which are, on average, more compact than the overall population. However, we found no differences in the mean sizes of our final sample of galaxies (both over-massive and under-massive), compared with the total population of galaxies with known black-hole masses (filled circles). Thus, our results are not driven by our rejection criteria in terms of signal-to-noise ratio or emission-line contamination.

doi:10.1038/nature25001

Large granulation cells on the surface of the giant star π^1 Gruis

C. Paladini[1,2], F. Baron[3], A. Jorissen[1], J.-B. Le Bouquin[4], B. Freytag[5], S. Van Eck[1], M. Wittkowski[6], J. Hron[7], A. Chiavassa[8], J.-P. Berger[4], C. Siopis[1], A. Mayer[7], G. Sadowski[1], K. Kravchenko[1], S. Shetye[1], F. Kerschbaum[7], J. Kluska[9] & S. Ramstedt[5]

Convection plays a major part in many astrophysical processes, including energy transport, pulsation, dynamos and winds on evolved stars, in dust clouds and on brown dwarfs[1,2]. Most of our knowledge about stellar convection has come from studying the Sun: about two million convective cells with typical sizes of around 2,000 kilometres across are present on the surface of the Sun[3]—a phenomenon known as granulation. But on the surfaces of giant and supergiant stars there should be only a few large (several tens of thousands of times larger than those on the Sun) convective cells[3], owing to low surface gravity. Deriving the characteristic properties of convection (such as granule size and contrast) for the most evolved giant and supergiant stars is challenging because their photospheres are obscured by dust, which partially masks the convective patterns[4]. These properties can be inferred from geometric model fitting[5–7], but this indirect method does not provide information about the physical origin of the convective cells[5–7]. Here we report interferometric images of the surface of the evolved giant star π^1 Gruis, of spectral type[8,9] S5,7. Our images show a nearly circular, dust-free atmosphere, which is very compact and only weakly affected by molecular opacity. We find that the stellar surface has a complex convective pattern with an average intensity contrast of 12 per cent, which increases towards shorter wavelengths. We derive a characteristic horizontal granule size of about 1.2×10^{11} metres, which corresponds to 27 per cent of the diameter of the star. Our measurements fall along the scaling relations between granule size, effective temperature and surface gravity that are predicted by simulations of stellar surface convection[10–12].

The giant star π^1 Gruis was observed with the four-telescope H-band beam combiner PIONIER[13] mounted at the Very Large Telescope Interferometer (VLTI; Cerro Paranal, Chile) during the nights of 2014 September 25 and 29. The observations (see also Methods section 'Data') cover three spectral channels across the near-infrared H band (central wavelengths,1.625 μm, 1.678 μm and 1.730 μm; width of the filters, 0.0483 μm, corresponding to a spectral resolution of 35). Owing to the excellent Fourier plane coverage acquired and to the good signal-to-noise ratio (Extended Data Fig. 1), we are able to reconstruct model-independent images in each spectral channel (Fig. 1). For this purpose, we use the image reconstruction software SQUEEZE[14], which is based on a Markov chain Monte Carlo (MCMC) approach to the regularized maximum likelihood problem. To assess the reliability of the image we also use a different image reconstruction algorithm, MiRa[15]. The images from SQUEEZE and MiRa have very similar characteristics (Fig. 1; see also Methods section 'Image reconstruction'). The dusty envelope that enshrouds the star is transparent in the wavelength range of our observations. The major molecular contributions in this wavelength range are CO and CN; this molecular contribution is the weakest in the longest-wavelength filter, which can be considered as probing the continuum. The images show a stellar disk with the diameter weakly

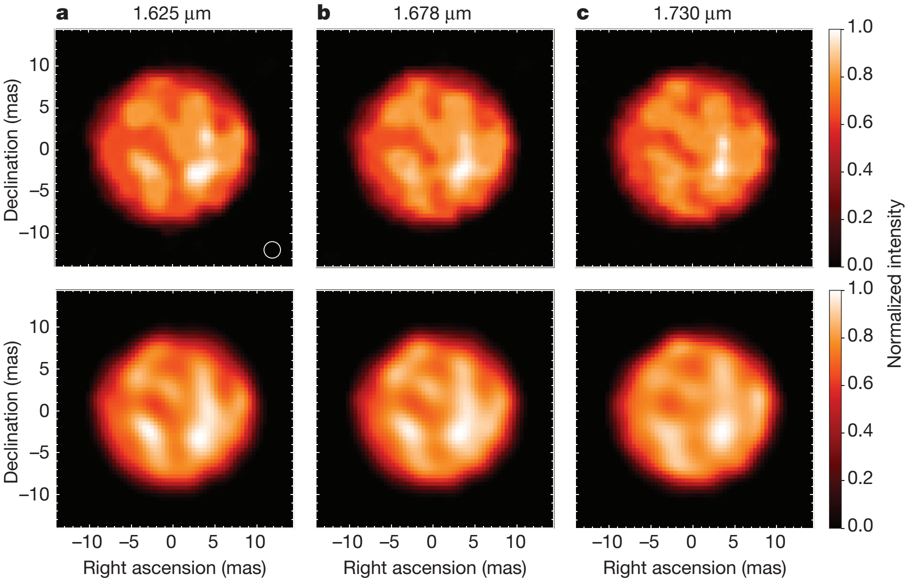


Figure 1 | The stellar surface of π^1 Gruis. **a**–**c**, Images of the stellar surface of π^1 Gruis reconstructed from the interferometric data using the SQUEEZE[14] algorithm (upper panels) or the MiRa[15] algorithm (lower panels). Images are shown in the spectral channels centred on 1.625 μm (**a**), 1.678 μm (**b**) and 1.730 μm (**c**). The angular resolution of the observations is $\lambda/(2B) \approx 2$ mas, where λ is the wavelength and B is the baseline, that is, the distance between two apertures, and is represented by the circle at the bottom right of the top panel in **a**. In each image, one pixel corresponds to 0.45 mas.

[1] Institut d'Astronomie et d'Astrophysique, Université libre de Bruxelles, CP 226, 1050 Bruxelles, Belgium. [2] European Southern Observatory, Alonso de Cordova 3107, Vitacura, Santiago, Chile. [3] Department of Physics and Astronomy, Georgia State University, PO Box 5060 Atlanta, Georgia 30302-5060, USA. [4] Université Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France. [5] Department of Physics and Astronomy, Uppsala University, Box 516, 75120 Uppsala, Sweden. [6] European Southern Observatory, Karl-Schwarzschild-Strasse 2, 85748 Garching bei München, Germany. [7] Department of Astrophysics, University of Vienna, Türkenschanzstrasse 17, 1180 Vienna, Austria. [8] Laboratoire Lagrange, UMR 7293, Université de Nice Sophia-Antipolis, CNRS, Observatoire de la Côte d'Azur, BP 4229, 06304 Nice Cedex 4, France. [9] University of Exeter, Department of Physics and Astronomy, Stocker Road, Exeter EX4 4QL, UK.

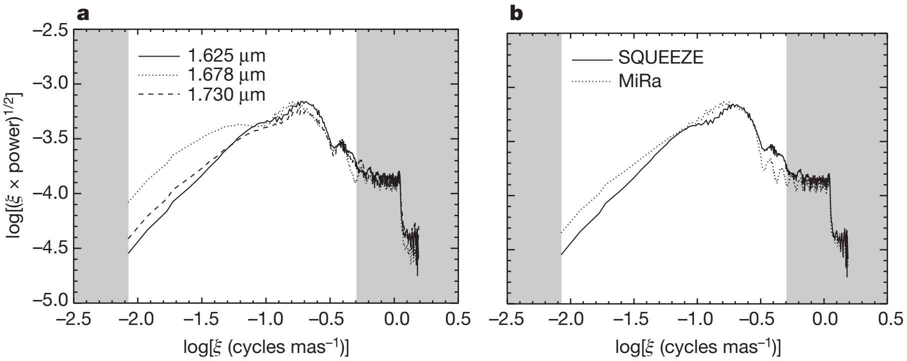


Figure 2 | Power spectral density. a, The resulting power spectrum (multiplied by ξ, where ξ is the number of wavelengths per unit of distance) from the SQUEEZE images in three different PIONIER spectral channels (see legend). The granule size is derived by averaging the maximum of the three curves after smoothing. **b**, Comparison between the power spectrum of the SQUEEZE and MiRa images corresponding to the first spectral channel of PIONIER (1.625 μm). The peak of the MiRa images corresponds to a granule size consistent (within error bars) with that derived from the SQUEEZE image. In **a** and **b**, the grey shaded area on the left marks the limit of the box of the image and the grey shaded area on the right marks the limit of the angular resolution of our observations ($\lambda/(2B)$, as indicated in Fig. 1).

dependent on the wavelength, indicating that the molecular envelope is very shallow and that we are probing the stellar surface continuum. There are patchy structures, all well within the nearly circular stellar surface (in contrast to, for instance, the situation for the well studied[7,16] supergiant Betelgeuse). For all of these reasons, we infer that the patterns seen on the stellar surface are the actual convective granules. Arguably, this is the most detailed model-independent (see Methods section 'Image reconstruction') image of convective patterns on the surface of a giant star obtained so far.

We estimate the intensity contrast[17] $\Delta I_{\rm rms}/\langle I\rangle$ after correcting the intensity profile for the limb-darkening effect (see Methods section 'Power spectrum'). This correction ensures that the contrast is sensitive only to the convective pattern. We obtain 13.1% ± 0.2%, 12.3% ± 0.4% and 11.9% ± 0.4% in the three spectral channels (ordered by increasing wavelength), where the (statistical) uncertainty corresponds to the standard deviation of the contrast on the various images produced by the MCMC approach used within the SQUEEZE code (see Methods section 'Image reconstruction'). The contrast measurements show a trend towards higher values at shorter wavelengths, indicative of a stronger molecular contamination. Three-dimensional models[18,19] of red giants are still at an exploratory stage and do not cover the parameter space of π^1 Gruis. Such models predict a bolometric intensity contrast for the surface convective motions of around 20%; contrasts in the H band, as observed here, may be expected to be lower. Systematic errors (resulting from seeing, limited resolution, and so on) that are difficult to assess quantitatively are also expected to make the observed contrast lower. The above effects (and others) have been discussed in the context of direct imaging of solar granulation[11,17], but they are likely to reduce the contrast in the interferometric context considered here as well.

The prospect offered by the granule size in that respect is much better. Although in previous studies[5–7] this quantity was obtained via model fitting, often unconvincingly because of the high values of the resulting reduced χ^2, our interferometric model-independent images allow us to derive the granule size directly from a spatial power-spectrum density (PSD) analysis. This technique is applied routinely on theoretical model predictions[10–12] and is used here to derive the wavenumber ($k = 2\pi\xi$; where $\xi = 1/\lambda$ is the number of wavelengths per unit of distance, not to be confused with the wavelengths λ of the observations) that carries the maximum power, that is, the characteristic granule size. In Fig. 2a we show the PSD as a function of spatial frequency for the three spectral channels of the PIONIER SQUEEZE image cube. After smoothing the PSD of the three spectral channels to remove the 'wiggles', we derive the maximum of the PSD by averaging the position of the three peaks. The error is calculated as the standard deviation of these three maximum values. The resulting granule size is 5.3 ± 0.5 mas. The PSD from the MiRa images gives very similar results (Fig. 2b), and the granule size agrees within the error with that derived from the SQUEEZE images. It is not possible to compare this observed value directly with a global model atmosphere that matches π^1 Gruis because such a model does not exist. Clearly, stellar surface convection does not look the same across the entire Hertzsprung–Russell diagram in terms of, for instance, granule size. However, if stellar convection is governed by the same (magneto-)radiation-hydrodynamic processes across the Hertzsprung–Russell diagram, then the parametric formulae that relate the characteristic granule size to the stellar parameters, which are based on predictions from the mixing-length theory of convection[20] and from models for less-evolved stars[10–12], might be applicable to π^1 Gruis as well.

To perform this comparison between the spatial structure detected at the surface of π^1 Gruis and these parametric relations, we convert the typical angular size of the granules observed at the surface of π^1 Gruis into a linear size. Using the π^1 Gruis parallax of 6.13 ± 0.76 mas (ref. 21), we obtain a characteristic linear granule size of $x_g = (1.2 \pm 0.2) \times 10^{11}$ m. The comparison also requires knowledge of the atmospheric

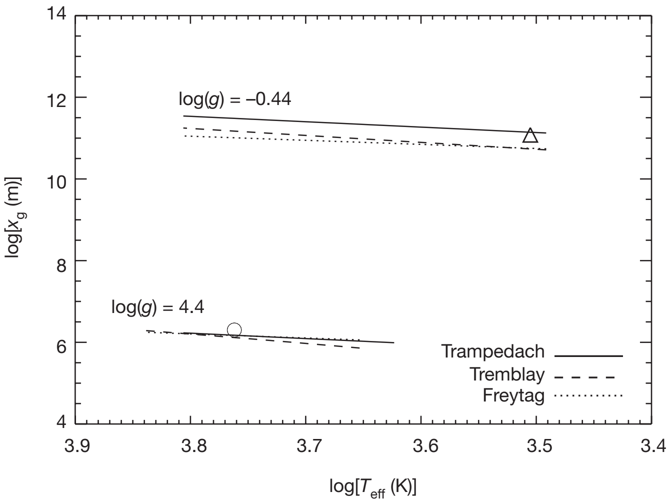


Figure 3 | The characteristic horizontal scale of granulation of π^1 Gruis versus theoretical predictions. The characteristic granule size x_g obtained from the PIONIER images of π^1 Gruis (triangle; $\log(g) = -0.44$) is quite different from the solar value (circle; $\log(g) = 4.4$). Even though theoretical predictions (dotted line[10], dashed line[11] and solid line[12]) had to be extrapolated to lower effective temperatures to match the stellar parameters of π^1 Gruis ($T_{\rm eff} = 3{,}200$ K, $\log(g) = -0.44$), they are in good agreement with the observations.

parameters of π^1 Gruis (see Methods section 'Stellar parameters'): effective temperature $T_{\rm eff} = 3{,}200$ K, surface gravity $\log(g) = -0.4$, solar metallicity and mean molecular weight $\mu = 1.3\,{\rm g\,mol^{-1}}$ for a non-ionized solar mixture with 75% hydrogen and 25% helium in mass.

The mixing-length theory of convection predicts that the characteristic granule size $x_{\rm g}$ scales linearly with the pressure scale height $H_{\rm p}$ (ref. 10):

$$x_{\rm g} = 10H_{\rm p} \tag{1}$$

where $H_{\rm p} = T_{\rm eff}R_{\rm gas}/(\mu g)$ and $R_{\rm gas} = 8.314 \times 10^7\ {\rm erg\,K^{-1}\,mol^{-1}}$ is the gas constant. The proportionality factor of 10 ensures that the formula correctly predicts the granulation at the surface of the Sun (ref. 11 and Fig. 3). Expressing $x_{\rm g}$ in units of 10^6 m, g in cm s^{-2} and $T_{\rm eff}$ in K, equation (1) becomes[10]

$$\log(x_{\rm g,Freytag}) = \log(T_{\rm eff}) - \log(g) - \log(\mu) + 0.92 \tag{2}$$

or $x_{\rm g,Freytag} = 5.1 \times 10^{10}$ m using the atmospheric parameters for π^1 Gruis.

More recent three-dimensional models[12] have extended the early analysis to FGK dwarfs and K giants, and predict that the size of the convective granules depends on the stellar parameters in a manner very similar to equation (2) (using the same units):

$$\log(x_{\rm g,Trampedach}) = (1.321 \pm 0.004)\log(T_{\rm eff}) - (1.0970 \pm 0.0003)\log(g) + (0.031 \pm 0.036)$$

This relation yields $x_{\rm g,Trampedach} = 1.2 \times 10^{11}$ m for π^1 Gruis. Finally, the CIFIST grid[11,22], which also covers M-type stars and sub-solar metallicities (expressed as $[{\rm Fe/H}] = \log[N({\rm Fe})/N({\rm H})]_{\rm star} - \log[N({\rm Fe})/N({\rm H})]_{\rm Sun}$, where N(A) is the number density of element A), follows[11]

$$\log(x_{\rm g,Tremblay}) = 1.75\log[T_{\rm eff} - 300\log(g)] - \log(g) + 0.05[{\rm Fe/H}] - 1.87$$

This relation yields $x_{\rm g,Tremblay} = 4.9 \times 10^{10}$ m for π^1 Gruis.

Despite the fact that none of the above relations is derived from stellar models that match the atmospheric parameters of π^1 Gruis (a very evolved late-type giant star), the predictions are in fairly good agreement (Fig. 3). This result suggests that predictions from current models of stellar surface convection may be extrapolated to the region of the Hertzsprung–Russell diagram where luminous red giants are located. In future studies, securing time-series images will enable the lifetime of the observed convective structures to be determined, providing another way of comparing observations with model predictions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 March; accepted 24 October 2017.

Published online 20 December 2017.

1. Giménez, Á. *et al.* (eds) *ASP Conf. Ser.* Vol. 173 (ASP, 1999).
2. Kupka, F. Convection in stars. In *Proc. IAU Symp.* Vol. 224 (eds Zverko, J. *et al.*) 119–129 (Cambridge Univ. Press, 2004).
3. Schwarzschild, M. On the scale of photospheric convection in red giants and supergiants. *Astrophys. J.* **195,** 137–144 (1975).
4. Wittkowski, M. *et al.* Aperture synthesis imaging of the carbon AGB star R Sculptoris: detection of a complex structure and a dominating spot on the stellar disk. *Astron. Astrophys.* **601,** A3 (2017).
5. Young, J. S. *et al.* New views of Betelgeuse: multi-wavelength surface imaging and implications for models of hotspot generation. *Mon. Not. R. Astron. Soc.* **315,** 635–645 (2000).
6. Haubois, X., Perrin, G. & Lacour, S. Imaging the spotty surface of Betelgeuse in the H band. *Astron. Astrophys.* **508,** 923–932 (2009).
7. Montargès, M. *et al.* The close circumstellar environment of Betelgeuse. IV. VLTI-PIONIER interferometric monitoring of the photosphere. *Astron. Astrophys.* **588,** A130 (2016).
8. Keenan, P. C. Classification of the S-type stars. *Astrophys. J.* **120,** 484–505 (1954).
9. Mayer, A. *et al.* Large-scale environments of binary AGB stars probed by Herschel. II. Two companions interacting with the wind of π^1 Gruis. *Astron. Astrophys.* **570,** A113 (2014).
10. Freytag, B., Holweger, H., Steffen, M. & Ludwig, H.-G. in *Science with the VLT Interferometer* (ed. Paresce, F.) 316–317 (Springer, 1997).
11. Tremblay, P.-E. *et al.* Granulation properties of giants, dwarfs, and white dwarfs from the CIFIST 3D model atmosphere grid. *Astron. Astrophys.* **557,** A7 (2013).
12. Trampedach, R. *et al.* A grid of three-dimensional stellar atmosphere models of solar metallicity. I. General properties, granulation, and atmospheric expansion. *Astrophys. J.* **769,** 18 (2013).
13. Le Bouquin, J.-B. *et al.* PIONIER: a 4-telescope visitor instrument at VLTI. *Astron. Astrophys.* **535,** A67 (2011).
14. Baron, F., Monnier, J. D. & Kloppenborg, B. A novel image reconstruction software for optical/infrared interferometry. *Proc. SPIE* **7734,** 77342I (2010).
15. Thiébaut, É. MIRA: an effective imaging algorithm for optical interferometry. *Proc. SPIE* **7013,** 70131I (2008).
16. Kervella, P. *et al.* The close circumstellar environment of Betelgeuse. II. Diffraction-limited spectro-imaging from 7.76 to 19.50μm with VLT/VISIR. *Astron. Astrophys.* **531,** A117 (2011).
17. Wedemeyer-Böhm, S. & Rouppe van der Voort, L. On the continuum intensity distribution of the solar photosphere. *Astron. Astrophys.* **503,** 225–239 (2009).
18. Freytag, B. & Höfner, S. Three-dimensional simulations of the atmosphere of an AGB star. *Astron. Astrophys.* **483,** 571–583 (2008).
19. Freytag, B., Liljegren, S. & Höfner, S. Global 3D radiation-hydrodynamics models of AGB stars. Effects of convection and radial pulsations on atmospheric structures. *Astron. Astrophys.* **600,** A137 (2017).
20. Ulrich, R. K. Convective energy transport in stellar atmospheres. I: a convective thermal model. *Astrophys. Space Sci.* **7,** 71–86 (1970).
21. van Leeuwen, F. Validation of the new Hipparcos reduction. *Astron. Astrophys.* **474,** 653–664 (2007).
22. Ludwig, H.-G. *et al.* The CIFIST 3D model atmosphere grid. *Mem. Soc. Astron. Ital.* **80,** 711–714 (2009).

Acknowledgements C.P. acknowledges the support of the Fonds National de la Recherche Scientifique (F.R.S.-FNRS), Belgium. C.S., G.S. are supported by the PRODEX office, Belgium. This research has been funded by the Belgian Science Policy Office under contract BR/143/A2/STARLAB (S.S., A.J., S.V.E.). K.K. is supported by a FRIA grant (Belgium). F.B. acknowledges funding by the National Science Foundation, NSF-AST numbers 1445935 and 1616483. J.K. acknowledges the Philip Leverhulme Prize (PLP-2013-110). S.V.E. thanks the Fondation ULB for its support. The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement 730890 (OPTICON) and from the Austrian Science Fund (FWF) under project AP23006-N16. We thank all the ESO/VLTI staff for supporting our observations.

Author Contributions This project was initiated by C.P. A.M. carried out the observations. J.B.L.B. performed the data reduction, F.B. and J.B.L.B. performed the image reconstruction. The contrast and the PSD analysis was carried on by C.P., B.F. and F.B. The stellar parameter determination was made by C.P., S.V.E., S.S., and K.K. A.J. gave crucial help through all the scientific analysis. The text was written by C.P., F.B. and A.J., and edited by the other authors that contributed to the scientific discussion.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to C.P. (cpaladin@eso.org).

Reviewer Information *Nature* thanks P. Kervella, J. Young and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data. The target of this study was observed with the PIONIER[13] instrument mounted at the Very Large Telescope Interferometer (VLTI; Cerro Paranal). The light from the four auxiliary telescopes, with 1.8-m apertures, was combined to obtain 303 spectrally dispersed visibilities and 201 closure phases (Extended Data Fig. 1). The maximum baseline used to collect the observations is 90 m, which corresponds to an array angular resolution $\lambda/(2B)$ of approximately 2 mas. The data reduction was performed using the *pndrs* package[13]. Following the standard procedure of PIONIER data reduction, the calibrated data are the average of five consecutive files of each observing block. A systematic relative uncertainty of 5% is added on the data. The stars HD 209688 (diameter[23], 2.62 ± 0.03 mas) and HD 215104 (diameter[24], 1.7 ± 0.1 mas) were used as calibrators. These objects were selected by using the SearchCal tool developed by the Jean-Marie Mariotti Center (JMMC).

Image reconstruction. The image reconstruction was performed using two different algorithms, SQUEEZE[14] and MiRa[15], to assess the reliability of the image reconstruction process. Using regularized maximum likelihood, MiRa minimizes a joint criterion which is the sum of (1) a likelihood term that measures the compatibility with the data and (2) a regularization term that imposes priors on the image. The relative weight between these two terms is controlled by the 'hyperparameter' factor μ, and we use the so-called L-curve approach to estimate its optimal value for each regularizer[25]. The output image is 64×64 pixels, with a pixel scale of 0.45 mas per pixel. After testing different priors and regularizations to identify possible spurious structures in the reconstructions, we concluded that a trustworthy image is obtained with the MiRa smoothness regularizer, without a prior image. SQUEEZE is based on an MCMC approach to the regularized maximum likelihood problem. It implements parallel simulated annealing and parallel tempering methods, enabling the use of non-convex or non-smooth regularizations that are not implemented in MiRa. SQUEEZE can also handle multi-wavelength datasets, again with the possibility of non-convex and non-smooth trans-spectral regularizations. The images of π^1 Gruis were reconstructed with the same pixel scale and field of view as used for the MiRA reconstruction.

We first ran 50 independent, parallel, simulated-annealing MCMC chains of 16×16-pixel reconstructions at a quarter of the final resolution (1.8 mas per pixel). We then ran 50 MCMC chains of 32×32-pixel images at half the final resolution (0.9 mas per pixel), initializing the chains using the mean image over the chains of the 16×16-pixel run. Finally, we ran 50 chains of 64×64-pixel reconstruction at full resolution (0.45 mas per pixel), initializing the chains with the mean image of the 32×32-pixel run. The final reconstruction is again the mean image over the 64×64-pixel chains. The whole procedure is designed to avoid falling into local minima, making sure the image is optimal at the lower resolutions before moving on to filling finer details.

Because selecting an adequate regularization is crucial for imaging quality, we determined an adequate combination of regularizers and regularizer weights by simulation. We generated 'ground truth' images of spotless and spotted limb-darkened disks, and using our code OIFITS-SIM (https://github.com/bkloppenborg/oifits-sim) we produced OIFITS data with the same (u, v) spatial frequency coverage and signal-to-noise ratio as the original π^1 Gruis dataset. We then followed the reconstruction procedure described above for several combinations of regularizers that are known to work for stellar surfaces (such as entropy, total variation, l_0 pseudo-norm, field-of-view centring and wavelets) and regularizer weights. We selected the 'optimal' combination and weights as those that minimized the absolute mean difference between the reconstructions and the ground-truth disks: the chosen combination and weights should not introduce features to the featureless disks and should recover existing spots well. We found that a combination of spot regularizer[26] and l_0 sparsity in the image plane give the best result.

The final SQUEEZE image has a reduced χ^2 of approximately 1.3 and is remarkably similar to the image reconstructed with MiRa. Images at one MCMC standard deviation from the mean display the same features overall, although the details of the bright spots differ slightly (Extended Data Fig. 2). We note that our reconstructions were performed on the three PIONIER spectral channels independently. For comparison, SQUEEZE was also used in polychromatic mode (using total variation as a trans-spectral regularizer), but we did not find any major differences between the polychromatic images and the channel-by-channel approach.

Power spectrum. The PSD is the integral over rings that comprise wavenumbers in a certain interval around $k = 2\pi\xi$, where $\xi = 1/\lambda$ is the number of wavelengths per unit of distance of the squared amplitude of the two-dimensional Fourier transform of the image[17]. The PSD analysis of the original images produces a dominant peak at the wavenumber that corresponds to the diameter of the star, followed by several lobes that contain information of higher order. To be able to separate the peak associated with the typical granule size from the one associated with the stellar diameter, we subtract the stellar disk from the image. To perform this step we designed two circular masks (one based on a Gaussian intensity profile and the other on the MARCS model that best fits the photometry) and a square mask fully enclosed on the stellar surface. The Gaussian and the MARCS intensity profiles both introduce some spurious signal in the final PSD, owing to the fact that the intensity profile does not match the reconstructed one well. In particular, the MARCS intensity profile is steeper than the observed one. The square mask provides a final image that is 24×24 pixels and is free of the limb effects. The square mask provides the cleanest PSD and is therefore the one discussed in the paper (Fig. 2).

Before proceeding with the PSD analysis, we increased the number of padding points and rescaled the background of the image to the mean intensity of the stellar surface. The method was also applied to the MiRa image, yielding a PSD very similar to the one derived from the SQUEEZE image, as shown in Fig. 2.

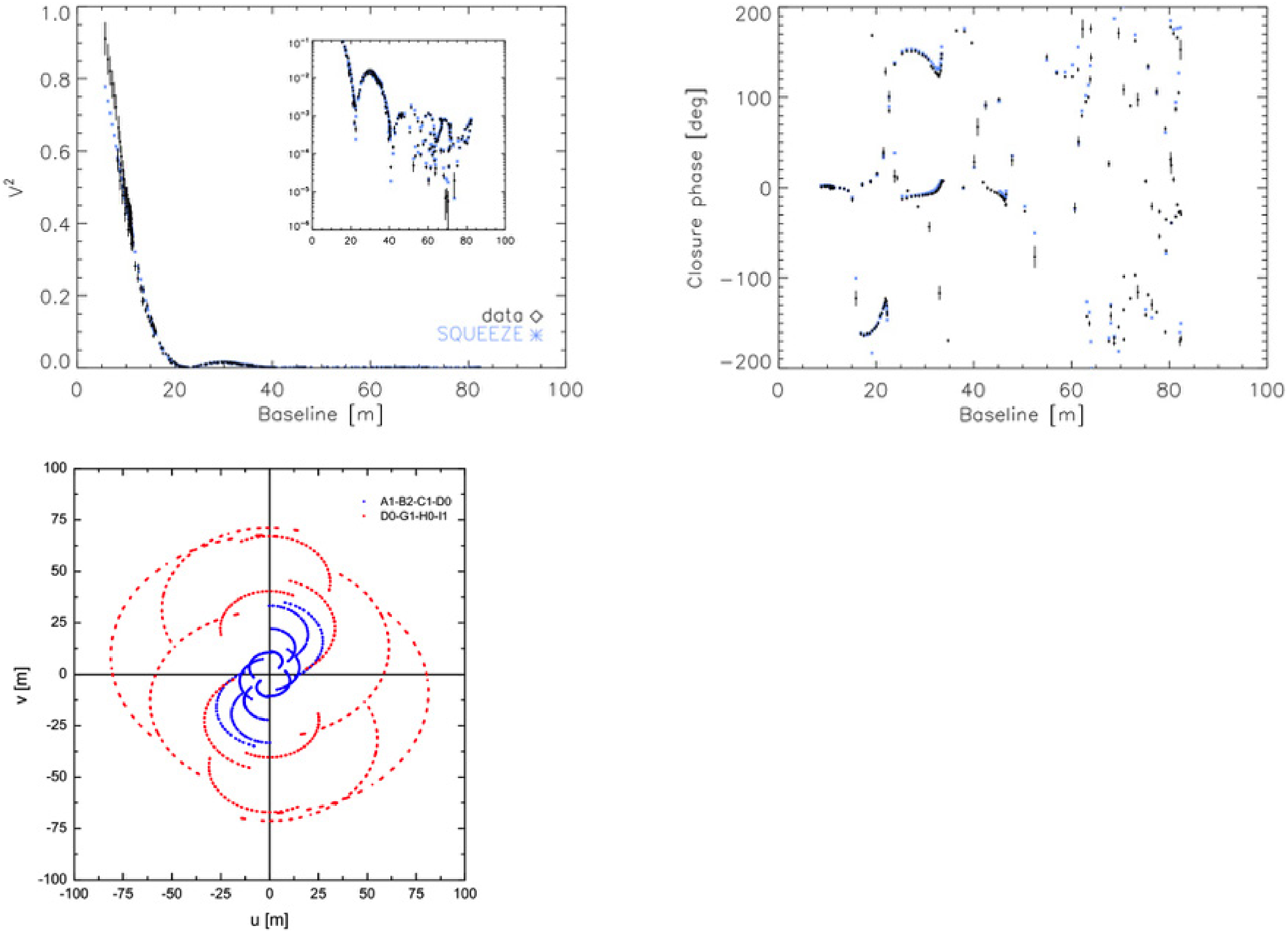
Stellar parameters. The stellar parameters of π^1 Gruis available from the literature[9] are the effective temperature T_{eff} of 3,100 K, a luminosity[9] of 7,200 times solar and a current mass[9] of 1.5 solar masses. Taking advantage of the new grid[27] of MARCS models with S-type chemistry, we decided to perform a new parameter estimate. On the basis of the literature values, we calculated a small grid of models that cover the following parameter space: $2{,}000\,\text{K} \le T_{\text{eff}} \le 3{,}200\,\text{K}$, with steps of 200 K; $\log(g) = 0$ and -0.44; 1, 1.3 and 2 solar masses; $[\text{s/Fe}] = 1$ dex and 2 dex; $\text{C/O} = 0.5$, 0.752, 0.899, 0.925, 0.951, 0.971 and 0.991; and a micro-turbulence of 2 km s^{-1}. The parameters of the model that best reproduce the photometry (Extended Data Fig. 3), which are later used for the comparison with the theoretical equations of the granule size, are $T_{\text{eff}} = 3{,}200$ K, $\log(g) = -0.44$, solar metallicity, s-process element abundances enhanced by 1 dex and $\text{C/O} = 0.991$. These values are in agreement with the literature ones. The diameter of π^1 Gruis was derived by fitting the PIONIER visibility data with a uniform disk. This choice is justified because intensity profiles derived from the MARCS model atmospheres are very similar to a uniform disk in the first visibility lobe. For the fitting we used the LitPro program provided by the JMMC.

The equivalent uniform disk diameter that we derived is 18.37 ± 0.18 mas, which corresponds to about 658 solar radii for a parallax[21] of 6.13 mas. Our diameter estimate is in agreement with literature values[9,28].

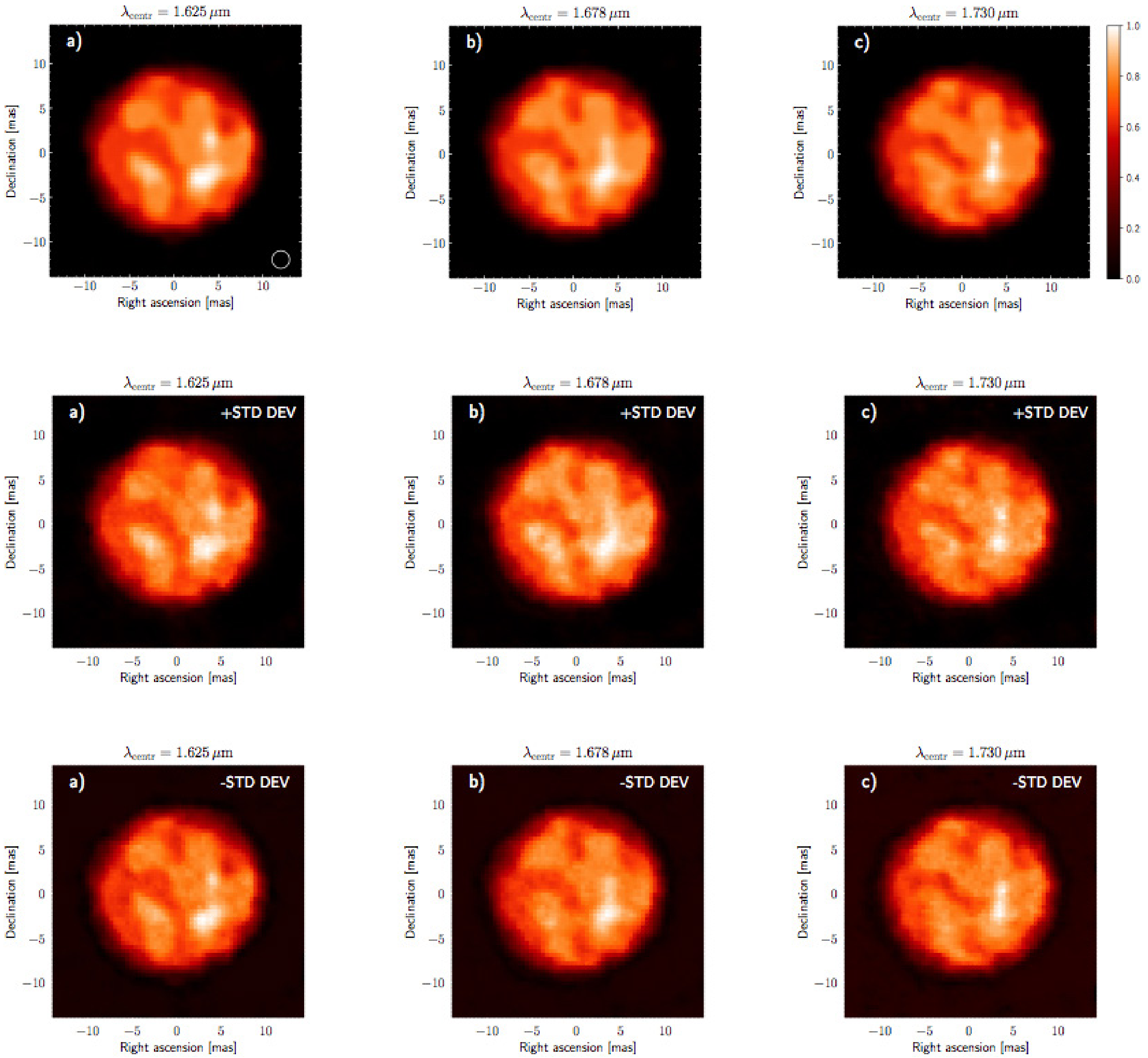
Code availability. The image reconstruction code SQUEEZE is publicly available at https://github.com/fabienbaron/squeeze. The image reconstruction code MiRa is publicly available at https://github.com/emmt/MiRA.

Data availability. The reduced PIONIER data sets are available in the http://oidb.jmmc.fr/ repository.

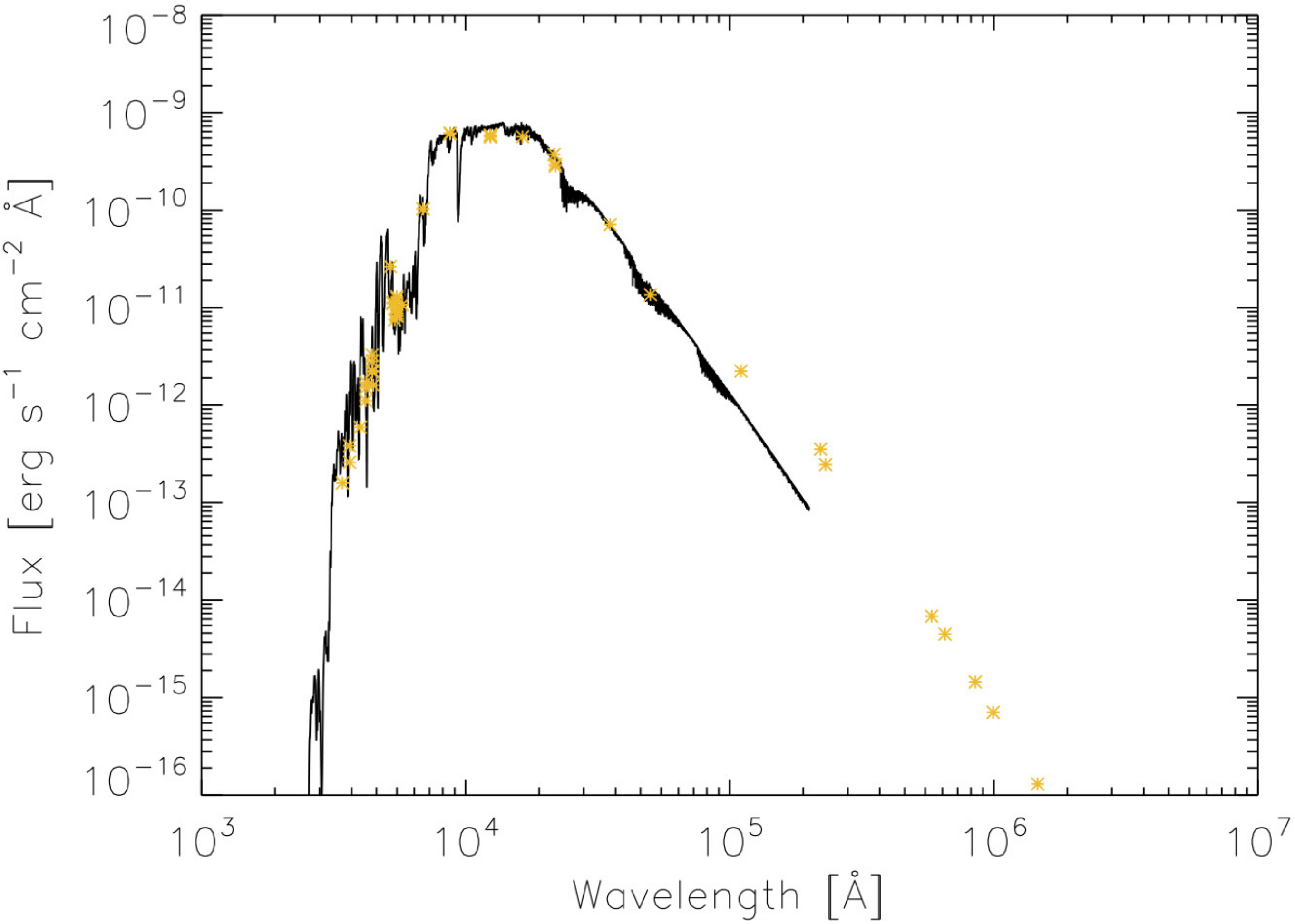
23. Bordé, P., Coudé du Foresto, V., Chagnon, G. & Perrin, G. A catalogue of calibrator stars for long baseline stellar interferometry. *Astron. Astrophys.* **393,** 183–193 (2002).
24. Lafrasse S. *et al.* Building the 'JMMC Stellar Diameters Catalog' using SearchCal. *Proc. SPIE* **7013,** 77344E11 (2010).
25. Renard, S., Thiébaut, E. & Malbet, F. Image reconstruction in optical interferometry: benchmarking the regularization. *Astron. Astrophys.* **533,** A64 (2011).
26. Baron, F., Monnier, J., Young, J. & Buscher, D. New theoretical frameworks for interferometric imaging. *ASP Conf. Ser.* **487,** 229–236 (2014).
27. Van Eck, S. *et al.* A grid of MARCS model atmospheres for late-type stars. II. S stars and their properties. *Astron. Astrophys.* **601,** A10 (2017).
28. Paladini, C. *et al.* The VLTI/MIDI view on the inner mass loss of evolved stars from the Herschel MESS sample. *Astron. Astrophys.* **600,** A136 (2017).



Extended Data Figure 1 | The PIONIER data. The upper left panel shows in black the PIONIER squared visibilities *V* and in blue the Fourier transform of the SQUEEZE image (first spectral channel) on a linear (main panel) and semi-logarithmic (inset) scale. The upper right panel is the same, but for the closure phase. The lower left panel shows the *u*–*v* coverage of the data. The blue data correspond to observations acquired with the short VLTI array configuration (called 'A1-B2-C1-D0'). The red data correspond to observations acquired with the intermediate array configuration ('D0-G1-H0-I1').



Extended Data Figure 2 | SQUEEZE images and error images. a–c, The first row of images corresponds to the adopted SQUEEZE images; the second row, labelled '+STD DEV', corresponds to images one standard deviation above the average image; and the third row of images, labelled '-STD DEV' shows images one standard deviation below the average image.



Extended Data Figure 3 | Spectral energy distribution. Comparison between the spectral energy distribution (yellow stars) and the best-fitting MARCS synthetic spectrum (black line). Note the presence of a moderate infrared excess longwards of 10 μm (10^{5} Å) attributable to circumstellar dust.

doi:10.1038/nature25137

Systems of mechanized and reactive droplets powered by multi-responsive surfactants

Zhijie Yang[1]*, Jingjing Wei[1]*, Yaroslav I. Sobolev[1] & Bartosz A. Grzybowski[1,2]

Although 'active' surfactants, which are responsive to individual external stimuli such as temperature[1], electric[2,3] or magnetic[4] fields, light[5,6], redox processes[6,7] or chemical agents[8], are well known, it would be interesting to combine several of these properties within one surfactant species. Such multi-responsive surfactants could provide ways of manipulating individual droplets and possibly assembling them into larger systems of dynamic reactors[9,10]. Here we describe surfactants based on functionalized nanoparticle dimers that combine all of these and several other characteristics. These surfactants and therefore the droplets that they cover are simultaneously addressable by magnetic, optical and electric fields. As a result, the surfactant-covered droplets can be assembled into various hierarchical structures, including dynamic ones, in which light powers the rapid rotation of the droplets. Such rotating droplets can transfer mechanical torques to their non-nearest neighbours, thus acting like systems of mechanical gears. Furthermore, droplets of different types can be merged by applying electric fields and, owing to interfacial jamming[11,12], can form complex, non-spherical, 'patchy' structures with different surface regions covered with different surfactants. In systems of droplets that carry different chemicals, combinations of multiple stimuli can be used to control the orientations of the droplets, inter-droplet transport, mixing of contents and, ultimately, sequences of chemical reactions. Overall, the multi-responsive active surfactants that we describe provide an unprecedented level of flexibility with which liquid droplets can be manipulated, assembled and reacted.

We developed surfactants based on dimeric nanoparticles comprising a smaller (6.0 ± 0.7 nm) gold domain and a larger (12.0 ± 1.2 nm) domain of either magnetic Fe_3O_4 or non-magnetic PbS (Fig. 1a). These particles were synthesized by using slightly modified literature

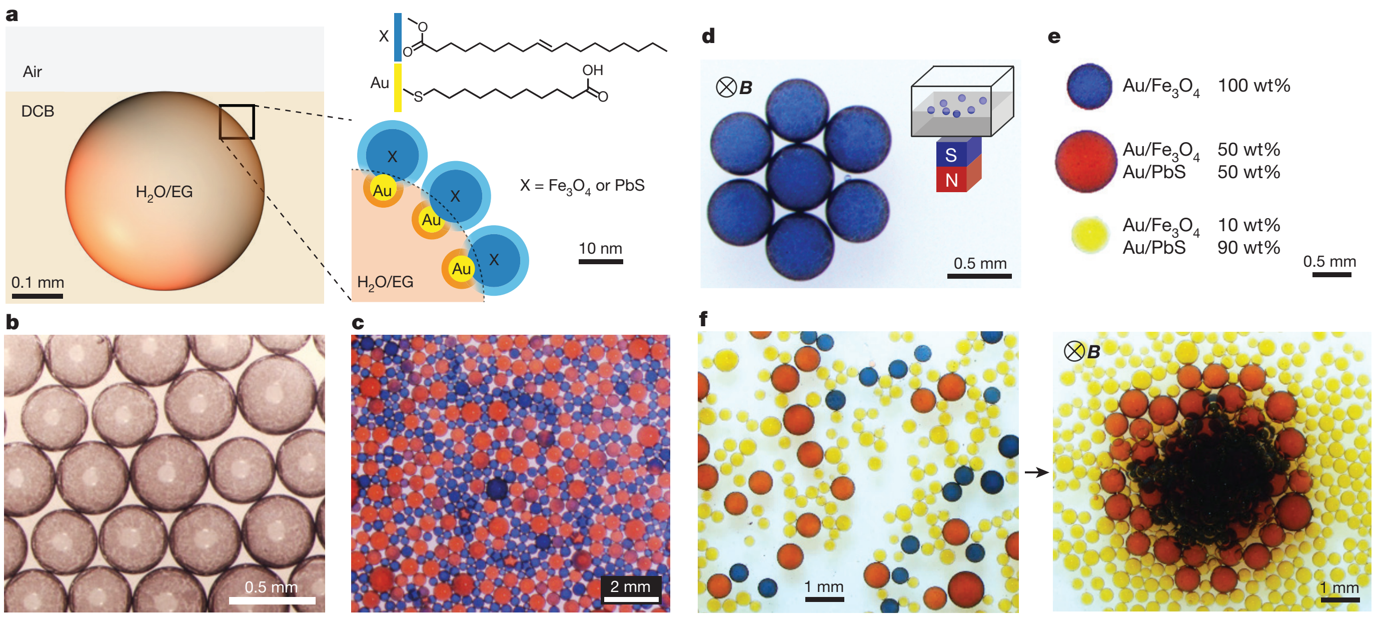


Figure 1 | Nanoparticle dimers as non-magnetic or magnetic surfactants. a, Schematic of a H_2O/EG droplet covered with MUA–Au/X–OA surfactants and suspended near the 1,2-DCB–air interface (EG, ethylene glycol; MUA, mercaptoundecanoic acid; X, diamagnetic PbS or magnetic Fe_3O_4; OA, oleic acid; DCB, dichlorobenzene). **b**, Optical micrographs of droplets coated with MUA–Au/Fe_3O_4–OA surfactants. The droplets are touching but neither merge nor exchange their contents. **c**, Two sub-populations of droplets are coloured with different dyes. The colours do not change for at least several days, further illustrating the lack of exchange of contents between droplets. **d**, A hexagonal structure formed by seven magnetic droplets in the field ***B*** produced by a 1.32 T permanent magnet (NdFeB block, K&J Magnetics D4Y0, magnetized along the long dimension) placed under the Petri dish that houses the droplets (see schematic in the inset). **e**, Images of droplets of different magnetic susceptibility, prepared by adjusting the proportions of magnetic MUA–Au/Fe_3O_4–OA and non-magnetic MUA–Au/PbS–OA surfactants. **f**, Hierarchical assembly of three types of droplet that differ in magnetic susceptibility. The NdFeB magnet is placed below the plane of the image. In **c**–**f**, the droplets (that is, the H_2O/EG phase) are coloured blue with methyl blue, red with Congo red and yellow with methyl orange, all at 0.3 g l^{-1}. See also Supplementary Video 1 for **d**–**f**.

[1]Center for Soft and Living Matter, Institute for Basic Science (IBS), Ulsan 44919, South Korea. [2]Department of Chemistry, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea.

*These authors contributed equally to this work.

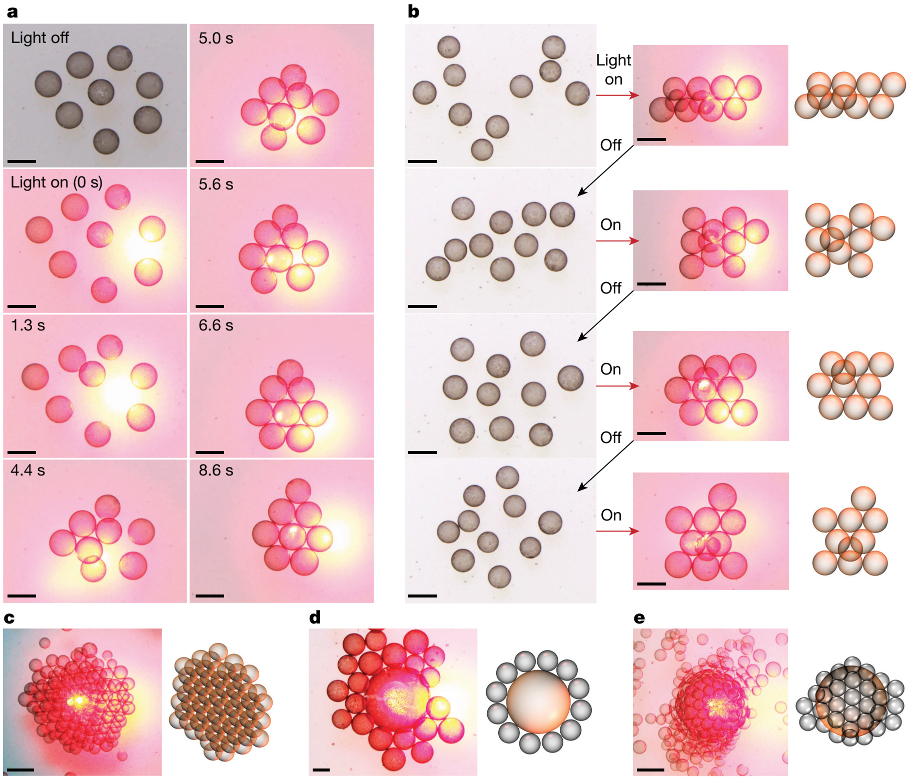


Figure 2 | Assembly of surfactant-stabilized droplets by using light. Droplets in all images were stabilized with MUA–Au/Fe_3O_4–OA surfactants, and manipulated and assembled using a diode laser (wavelength $\lambda_0 = 660 \pm 3$ nm, power $P_0 = 70 \pm 5$ mW; DL-7147-201, Tottori SANYO Electric). Unless otherwise noted, the droplets were suspended in a density-matched 4:3 v/v mixture of toluene and DCB ($\rho = 1.05$ g ml^{-1}). Instantaneous positions of the laser beam correspond to the bright-yellow regions. **a**, Individual frames from Supplementary Video 2. Irradiation of approximately 900 μm droplets causes them to assemble into a hexagonally close-packed, two-dimensional structure. Scale bars, 1 mm. **b**, Light-controlled assembly–disassembly of approximately 900 μm droplets into various two-layer structures (see Supplementary Video 2). The rightmost column shows schematics; all other images are from the experiment. The changes occur on a timescale of 10–20 s. Scale bars, 1 mm. **c**, Assembly of like-sized droplets into a three-dimensional close-packed structure (Supplementary Video 2). All droplets are about 500 μm in diameter; scale bar, 1 mm. **d**, Binary assembly of large (around 1,500 μm in diameter) and small (around 500 μm) droplets into a core–satellite structure (Supplementary Video 3). The density of both types of droplet is identical, $\rho = 1.05$ g ml^{-1}, and adjusted by the same EG volume fraction (50% v/v). **e**, Binary assembly of large (about 2,000 μm) and small (about 400 μm) droplets into a three-dimensional curvilinear structure (Supplementary Video 3). Here, the density of the small droplets ($\rho = 1.02$ g ml^{-1}, 25% v/v EG solution) is lower than that of the large droplet ($\rho = 1.05$ g ml^{-1}, 50% v/v EG solution). Scale bar, 1 mm. In **c**–**e**, the images on the left are from experiments and those on the right are schematics.

procedures[13,14] and were subsequently functionalized with two types of ligand: gold with 11-mercaptoundecanoic acid (MUA) terminated in a polar COOH head group (with an HLB index[15] of about 2.1 indicating hydrophilicity); and Fe_3O_4 or PbS with oleic acid (OA), which renders these domains hydrophobic. We denote these dimers MUA–Au/Fe_3O_4–OA and MUA–Au/PbS–OA. Unlike single-component particles covered with either MUA or OA, the dimers thus made were amphiphilic and acted as surfactants, forming monolayers at interfaces between various aqueous and organic phases (Supplementary Fig. 1). In most of our experiments, we used 400–2,000 μm droplets of a 1:1 mixture of water and ethylene glycol (H_2O/EG), suspended in either 1,2-dichlorobenzene (DCB) or a density-matched mixture of toluene and DCB (density $\rho = 1.05$ g ml^{-1}). In these solvent systems, the droplets did not raise and spread over the surface of the organic phase, but instead were buoyant and localized slightly below this surface (see schematic in Fig. 1a). As illustrated in Fig. 1b, c, the H_2O/EG droplets covered with the dimeric-nanoparticle surfactants did not mix with each other and did not exchange their contents over at least several days. For further experimental details, see Supplementary Information section 1.

When stabilized with MUA–Au/Fe_3O_4–OA surfactants (magnetic susceptibility[16] $\chi_1 = 5.3 \times 10^3$), the droplets responded to and followed external magnets, and remained intact when forming ordered assemblies in high-field regions (Fig. 1d, Supplementary Video 1). Their effective magnetic susceptibility could be modulated by covering the droplets with mixtures of magnetic (MUA–Au/Fe_3O_4–OA) and non-magnetic (MUA–Au/PbS–OA) surfactants in various proportions. In Fig. 1e we show three such droplets, with the liquid inside coloured using different dyes (methyl blue, Congo red or methyl orange) for visualization: blue droplets are covered with only MUA–Au/Fe_3O_4–OA particles, red droplets are covered with a 1:1 mass ratio of MUA–Au/Fe_3O_4–OA and MUA–Au/PbS–OA, and yellow droplets are covered with a 1:9 mass ratio of these surfactants. When a permanent magnet is placed beneath the vessel that houses a mixture of these droplets, they organize into a hierarchical structure with the most magnetic blue droplets at the centre, followed by the intermediate-susceptibility red droplets around them, and the least susceptible yellow droplets mostly at the periphery (Fig. 1f, Supplementary Video 1).

The droplets could also be manipulated using light absorbed by the nanoparticles (here, from a 660 nm diode laser; see Supplementary Videos 2 and 3). Upon irradiation, the droplets move towards the laser beam and within seconds assemble into various close-packed structures, such as the one shown in Fig. 2a. When the laser is switched off, the droplets disassemble on a timescale of about 20 s, but can be reassembled by turning the laser beam back on (Fig. 2b). This assembly–disassembly cycle can be performed multiple times (we tried hundreds) without any noticeable difference in droplet behaviour. In the assembled structures, the neutrally buoyant droplets can be

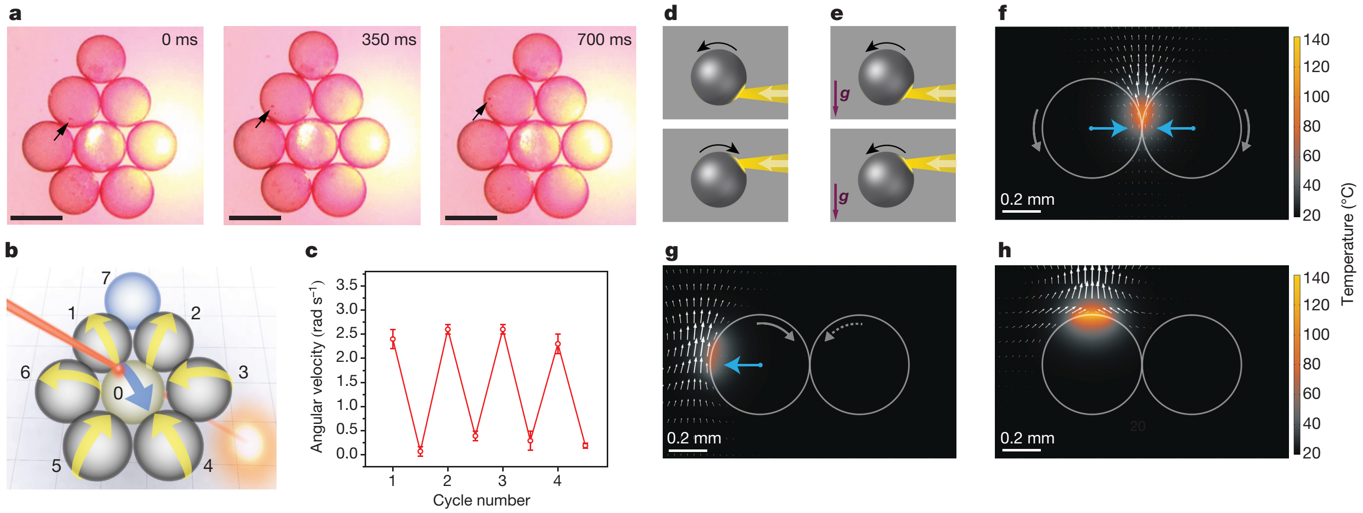


Figure 3 | Mechanized assemblies of rotating droplets. a, **b**, An assembly of eight droplets, with the laser beam focused at the edge of the central droplet (denoted as 0 in **b**). The motions of the droplets are clearly visible in Supplementary Video 4 and in the experimental images in **a**, and can be visualized by the motion of tracers at the surface (indicated by black arrows). In the schematic in **b**, the direction of these motions for droplets 1–6 are indicated by curved arrows; droplet 7 does not experience systematic directional motion (see main text for details). **c**, When the laser beam is focused at the centre of droplet 0, all droplets stop rotating. The motion resumes when the beam is moved back to the edge. These changes are quantified in the plot, which shows the average velocity of droplets 1–6, powered by droplet 0 being irradiated at the edge (droplets rotate) or at the centre (droplets stop rotating). Error bars correspond to standard deviations from five measurements. **d**, **e**, Two possible scenarios for light-induced rotations. If surface tension were dominant, changing the focal point of the laser would change the rotation direction of the droplets (**d**). However, we observe that the droplets rotate in the same direction irrespective of the position of the laser (**e**; the vector ***g*** indicates the direction of gravity). Calculated flows can be seen in Supplementary Video 13. **f**–**h**, Simulations (0.1 s after the beginning of heating) of convection in a system of two droplets heated from above by a laser beam directed at: the contact point between the droplets (**f**), the outer side of one of the droplets (**g**) or the top of one of the droplets (**h**). The cross-sections shown pass through the centres of the droplets and are parallel to the gravity vector. Colour represents temperature. White arrows show flow directions of toluene/DCB and speeds; the length of the arrows is proportional to velocity. Curved grey arrows indicate the direction of rotation of the droplets; the dashed arrowtail indicates that the right droplet is rotated by the left one via contact, not by drag due to convection. Blue arrows denote net force(s) acting on the droplets. Forces and torques are negligible in **h**. The maximum flow speed is approximately 0.9 mm s^{-1} in **f** and about 0.6 mm s^{-1} in **g** and **h**. See Fig. 5a and Supplementary Video 10 for visualizations of flows inside droplets. In all simulations, laser power is 70 mW, and 30% of it is absorbed and converted to heat by the surface of the droplet. See Supplementary Information section 2F for further theoretical details.

positioned within more than one layer, as in Fig. 2b in which the ones in the upper layer (that is, closer to the surface of the organic phase) are hexagonally close-packed and those in the lower layer are positioned at the triangular vacancies of the lattice of droplets above. With smaller droplets, such three-dimensional organization is even more pronounced, and closely packed, multilayer structures such as those in Fig. 2c are observed. We note that the spatial extent of these assemblies is limited to the region of irradiation. In Fig. 2c, the droplets within the approximately 5 mm diameter of the laser beam are held together tightly whereas those at the periphery are bound only loosely, attaching and detaching from the aggregate. Finally, when the droplets differ in size (Fig. 2d), the largest one localizes at the centre and is surrounded by a halo of smaller droplets. When the smaller droplets are also made to be lighter than the larger ones (by adjusting the EG content), they can 'climb' onto and organize over the large, central droplet (Fig. 2e).

An interesting variant of light-induced motion is observed when the laser beam is incident only at the edge of a droplet. In such a case, within about 100 ms from the onset of irradiation, the droplet starts to rotate rapidly. This effect is illustrated by experiments shown in Fig. 3a and Supplementary Video 4. Eight 900 μm droplets are first assembled into a compact structure by an unfocused laser beam, as in Fig. 2. The beam is then focused at the edge of the central droplet (denoted as '0' in the schematic in Fig. 3b). This droplet rotates at around 4 rad s^{-1} and induces rotational motion of the surrounding droplets (labelled 1–6), the directions of which are opposite to that of the powering droplet 0 (see curved arrows in Fig. 3b). The rotation of these nearest-neighbour droplets is slightly slower, at around 2.5 rad s^{-1}. Droplet 7 in Fig. 3b does not exhibit systematic rotation because the neighbouring droplets (1 and 2) try to rotate it in opposite directions—making an analogy to mechanical gears transmitting torques, gears 1 and 2 jam gear 7. This analogy is further corroborated by Supplementary Video 5 and Supplementary Fig. 15, in which the droplets are arranged so that the torque can be transmitted efficiently to non-nearest neighbours. The direction of rotation can be changed by changing the position of the focused beam around the perimeter of the droplet. The one exception here is when the beam is focused at the very centre of the droplet, in which case there is no rotation (Fig. 3c). We note that torque transmission between contacting droplet 'gears', even for long rotation times, does not result in the contents of the droplets being exchanged, as illustrated by experiments with dye-loaded droplets (Supplementary Video 5). Such behaviour is in sharp contrast to that of droplets that are stabilized by conventional molecular surfactants, which, as we and others[17,18] observed, merge readily under applied shear.

The origin of the motions of the droplets is explained by fluid-mechanical simulations (Supplementary Information section 2). In brief, the surfactant layer absorbs about 30% of the incident laser light and within 10–100 ms the temperature at the focal spot reaches the boiling points of both water and toluene (see Supplementary Video 6 for a visualization of the cavitation of small vapour bubbles). Such localized heating can change the surface tension and set up convective flows. However, surface tension effects do not depend on the direction of gravity: if they were dominant, we would expect the droplet to turn in opposite directions when heated over its upper and lower portions (Fig. 3d). In reality, we observed that the droplet always turns in the same direction (Fig. 3e, Supplementary Video 7). This phenomenon is due to the laser-heated fluid convecting upwards, dragging the surface of the droplet along and ultimately causing the droplet to always rotate away from the source of light or heat. Interestingly, when the droplet is heated over its upper part, the convective flows also give rise to a horizontal force that attracts the droplet towards the laser beam

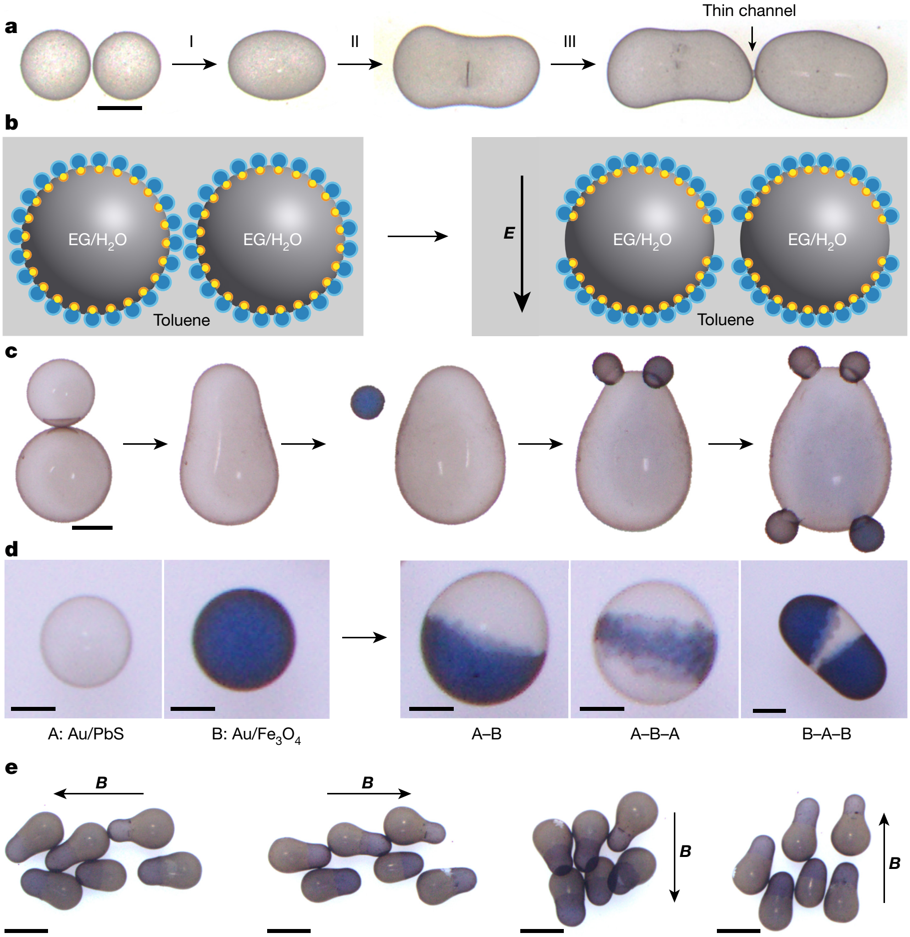


Figure 4 | Electrostatic 'welding' of droplets with complex shapes and of patchy droplets. a, The three sequential stages of welding. The arrow points to a thin channel joining the two domains in the structure on the right. Scale bar, 0.5 mm. **b**, The mechanism of droplet welding, in which the surfactants redistribute under a field-induced, dielectrophoretic force. **c**, Welding of droplets with complex shapes, consisting of a transparent 'trunk' and four dye-coloured 'arm' droplets. Scale bar, 0.5 mm. **d**, Formation of various magnetic and non-magnetic patchy droplets. The positioning of the non-magnetic (Au/PbS; A, clear), and magnetic (Au/Fe_3O_4; B, blue) surface domains is dictated by the number and the placement of individual droplets being coalesced. Scale bars, 0.5 mm. **e**, In the pear-shaped droplets, made by welding larger (around 1,000 μm in diameter) non-magnetic droplets and smaller (around 500 μm) magnetic ones, the dipoles are in the plane of the image; when an external magnet is moved around the dish, the droplets follow it like little compasses (see also Supplementary Fig. 19b, c, Supplementary Video 9). Scale bars, 1 mm.

(which explains the assembly of the droplets in Fig. 2). In contrast, when its bottom part is heated, the droplet is cyclically repelled from and then attracted to the light spot (Supplementary Video 8). When the droplet is heated at its top centre, the convective flows are symmetric and neither rotation nor translation are observed (Fig. 3f–h).

Whereas magnetic and optical fields can be used to move and position the droplets, electric fields enable them to be 'welded' into non-spherical structures. Fig. 4a illustrates a sequence of experiments in which short pulses of an electric field produced by a Zerostat gun are applied (about 10 kV near the tip of the gun). First, two proximal, approximately 500 μm droplets are merged into an oval-shaped droplet (stage I); then, two such ovals are merged into a dumbbell-shaped droplet (stage II); and finally, two dumbbells coalesce into a structure with two domains connected by a thin, approximately 200 μm channel (stage III). All of these non-spherical shapes persist for at least two days, which can be explained by interfacial jamming of the surfactants[11,12]. The mechanism of the welding can be attributed to the redistribution of the surfactant particles. Specifically, because the dielectric constant of the droplet ($\varepsilon = 65$ for a 1:1 v/v H_2O/EG mixture) is much higher than that of the surrounding liquid ($\varepsilon = 9.9$ for DCB), the external field induces a dielectrophoretic force under which the particles migrate towards the top and bottom poles of the droplet[19], in effect reducing the particle density along the equator and allowing the formation of liquid bridges between proximal drops (Fig. 4b). The welding can be used flexibly to create droplets of complex shapes (Fig. 4c) and comprising different types of surface domain[20–24]. The latter is illustrated in Fig. 4d, in which differently coloured regions are covered with different (magnetic or non-magnetic) surfactants. Such 'Janus' or 'patchy' droplets can act as, for instance, magnetic dipoles that orient (Fig. 4e, Supplementary Fig. 19b–d, Supplementary Video 9) or flip (Supplementary Fig. 19a, Supplementary Video 9) in an external magnetic field.

All of the above control modalities are combined in droplet systems that support chemical reactions. Fig. 5a and Supplementary Video 10 illustrate the important ability to mix the contents of merging droplets efficiently. In this example, a dumbbell 'reactor' is made by electrostatically welding two spherical droplets (containing viscous solutions of Congo red or HCl) and is kept stationary by a magnet positioned below the dish. Irradiation with a laser beam causes local heating of the surfactants, sets up convective flows inside the reactor and results in rapid mixing (in a few minutes, in contrast to hours without light). In Fig. 5b and Supplementary Video 11, the droplets that carry $CoCl_2$ and 2-methylimidazole are brought into contact by laser light, welded by electrostatic discharge and mixed by light to produce

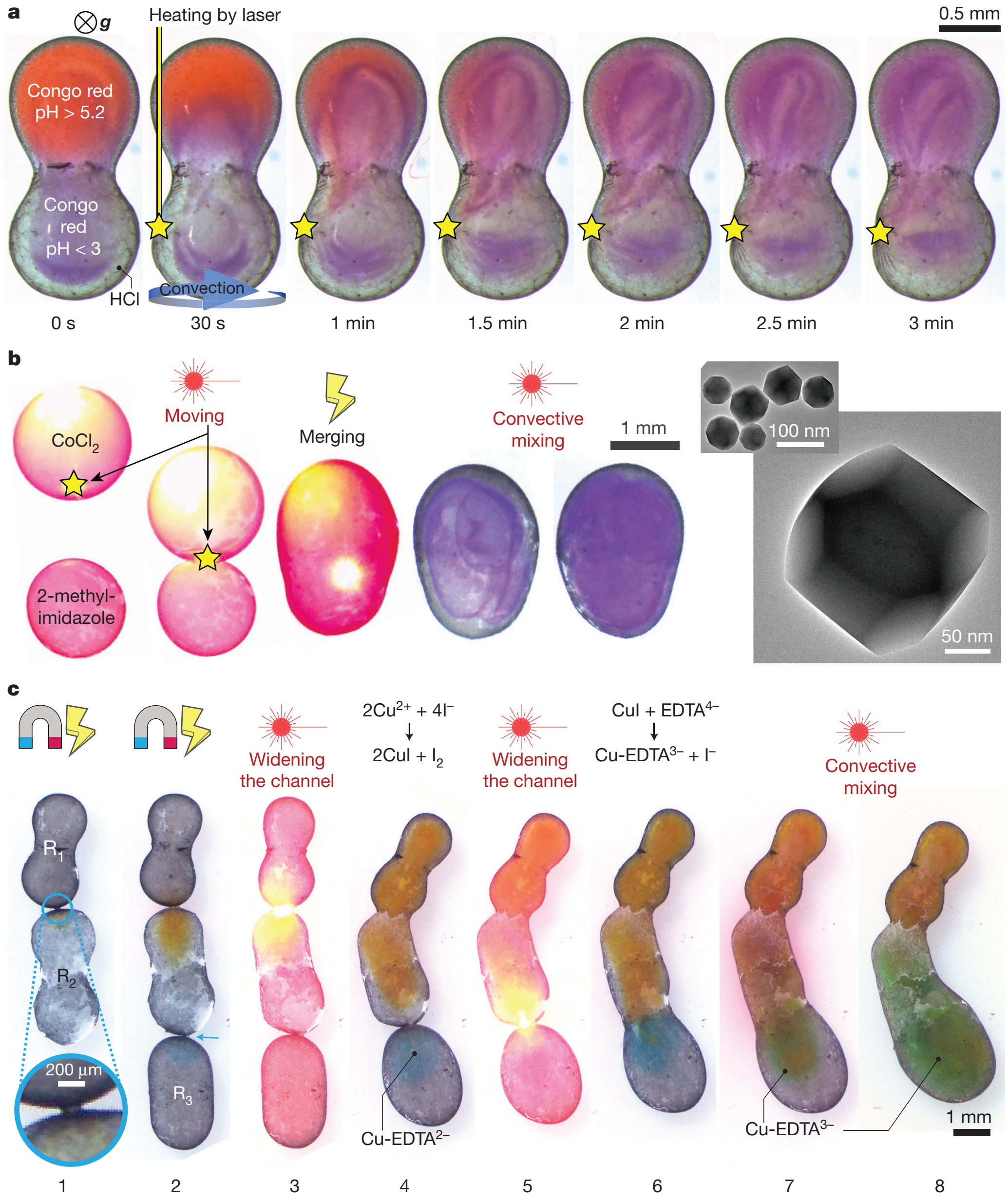


Figure 5 | Mixing and chemical reactions in dynamically controlled droplet 'reactors'. In all examples, the droplets are covered with MUA–Au/Fe_3O_4–OA magnetic surfactants. **a**, Convective mixing inside a dumbbell-shaped reactor, made by electrostatically welding a droplet containing 0.4 mM solution of Congo red in a 1:1 v/v mixture of EG/H_2O and a droplet containing 1.8 mM HCl, also in a 1:1 v/v mixture of EG/H_2O. The dumbbell is kept stationary in DCB by a permanent magnet that is placed below the plane of the image. A 660 nm laser beam (diameter, 0.14 mm; modulation frequency, 20 Hz; duty cycle, 50%; average power density, 200 W cm^{-2}) irradiates the spot marked by a yellow star and sets up convective flows. These flows are visualized by the violet colour, which corresponds to acidified Congo red. The gravity vector $\boldsymbol{g}$ points into the page. The curved blue arrow denotes the general direction of convective flow in the half of the dumbbell that is irradiated; flow in the other half is more complicated (see Supplementary Video 10). **b**, A laser (red laser symbol) guides the approach of an approximately 1.5 mm droplet containing 0.085 M $CoCl_2$ in water and an approximately 1.2 mm droplet containing 3.3 M 2-methylimidazole in water. Both droplets are suspended in DCB in a PTFE Petri dish. Upon contact and electrostatic welding (lightning bolt symbol), the droplets merge. Laser light is used to accelerate mixing. Eventually, after about 2 min, the reaction produces microcrystals of a ZIF-67 metal–organic framework. Transmission electron microscopy images of these crystals are shown on the far right. See also Supplementary Video 11. **c**, A sequence of reactions, $2Cu^{2+} + 4I^- \rightarrow 2CuI + I_2$ and $CuI + EDTA^{4-} \rightarrow Cu\text{-}EDTA^{3-} + I^-$. All three dumbbell-shaped reactors are filled with 1:1 v/v mixture of EG/H_2O. Reactor R_1 carries 1 μmol $CuSO_4$; R_2, 2 μmol KI; and R_3, 1 μmol Na_4EDTA. Reactors are oriented by the external magnetic field (magnet symbol; which is kept on for all frames 1–8 to keep the reactors in place) and electrostatically welded to connect via channels approximately 200 μm wide (steps 1 and 2; connecting channel shown in the inset of 1). The channels are widened and the contents mixed by laser light (R_1–R_2 in steps 3 and 4; R_2–R_3 in steps 5–7). Yellow-brown colour is due to I_2 produced in the first reaction; green is due to the mixing of I_2 with the $Cu\text{-}EDTA^{3-}$ complex produced in the second reaction. See Supplementary Video 12.

microcrystals of zeolitic imidazole framework 67 (ZIF-67)[25]. Finally, Fig. 5c and Supplementary Video 12 illustrate an appropriately timed sequence of two reactions that span three droplet 'reactors', denoted R_1, R_2 and R_3. Owing to their dumbbell shape, the reactors orient along an external magnetic field; this is important because for such shapes electrostatic welding opens only narrow (about 200 μm; see Fig. 4c) channels, through which the contents diffuse very slowly. However, when these narrow junctions are heated by the laser, they widen, allowing much faster transport and also rapid convective mixing (see above). In the system shown, light first widens the R_1–R_2 junction, enabling the reaction $2Cu^{2+} + 4I^- \rightarrow 2CuI + I_2$. Subsequently, the R_2–R_3 junction is widened, allowing CuI to react with $ETDA^{4-}$ to yield the final product of the sequence, $Cu\text{-}EDTA^{3-}$.

Looking forward, because systems of dynamic droplet reactors[26–29] are easy to make, manipulate, mix and 'valve', they can be complementary to microfluidic circuits that require more sophisticated fabrication and that can be limited to only certain types of solvent. We also see potential uses for the responsive surfactant droplets in studies of artificial cells and, with droplets containing curable gels or polymers, as light-guided inks from which complex shapes could be first assembled and then solidified.

Data Availability The data that support the findings of this study are available from the corresponding author on request. Project files used for fluid dynamics simulations are available in a public GitHub repository (https://doi.org/10.5281/zenodo.1040169).

Received 12 June; accepted 8 November 2017.
Published online 10 January 2018.

1. Meguro, K., Ueno, M. & Eumi, K. *Nonionic Surfactants: Physical Chemistry* 927–970 (Marcel Dekker, 1987).
2. Ha, J. & Yang, S. Effect of nonionic surfactant on the deformation and breakup of a drop in an electric field. *J. Colloid Interface Sci*. **206,** 195–204 (1998).
3. Prasad, M., Stubbe, F., Beunis, F. & Neyts, K. Different types of charged-inverse micelles in nonpolar media. *Langmuir* **32,** 5796–5801 (2016).
4. Brown, P. *et al.* Magnetic control over liquid surface properties with responsive surfactants. *Angew. Chem. Int. Ed*. **51,** 2414–2416 (2012).
5. Eastoe, J. & Vesperinas, A. Self-assembly of light-sensitive surfactants. *Soft Matter* **1,** 338–347 (2005).
6. Rosslee, C. & Abbott, N. L. Active control of interfacial properties. *Curr. Opin. Colloid Interface Sci*. **5,** 81–87 (2000).
7. Gallardo, B. S. *et al.* Electrochemical principles for active control of liquids on submillimeter scales. *Science* **283,** 57–60 (1999).
8. Liu, Y., Jessop, P. G., Cunningham, M., Eckert, C. A. & Liotta, C. L. Switchable surfactants. *Science* **313,** 958–960 (2006).
9. Maglia, G. *et al.* Droplet networks with incorporated protein diodes show collective properties. *Nat. Nanotechnol*. **4,** 437–440 (2009).
10. Villar, G., Graham, A. D. & Bayley, H. A tissue-like printed material. *Science* **340,** 48–52 (2013).
11. Cui, M., Emrick, T. & Russell, T. Stabilizing liquid drops in nonequilibrium shapes by the interfacial jamming of nanoparticles. *Science* **342,** 460–463 (2013).
12. Subramanian, A. B., Abkarian, M., Mahadevan, L. & Stone, H. A. Colloid science: non-spherical bubbles. *Nature* **438,** 930 (2005).
13. Wu, B., Zhang, H., Chen, C., Lin, S. & Zheng, N. Interfacial activation of catalytically inert Au (6.7 nm)-Fe_3O_4 dumbbell nanoparticles for CO oxidation. *Nano Res*. **2,** 975–983 (2009).
14. Shi, W. *et al.* A general approach to binary and ternary hybrid nanocrystals. *Nano Lett*. **6,** 875–881 (2006).
15. Griffin, W. C. Classification of surface-active agents by "HLB". *J. Soc. Cosmet. Chem*. **1,** 311–326 (1949).
16. Yu, H. *et al.* Dumbbell-like bifunctional Au−Fe_3O_4 nanoparticles. *Nano Lett*. **5,** 379–382 (2005).
17. Vanapalli, S. A. & Coupland, J. N. Emulsions under shear—the formation and properties of partially coalesced lipid structures. *Food Hydrocoll*. **15,** 507–512 (2001).
18. Shardt, O., Derksen, J. J. & Mitra, S. K. Simulations of droplet coalescence in simple shear flow. *Langmuir* **29,** 6201–6212 (2013).
19. Nudurupati, S., Janjua, M., Singh, P. & Aubry, N. Effect of parameters on redistribution and removal of particles from drop surfaces. *Soft Matter* **6,** 1157–1169 (2010).
20. Jiang, S. *et al.* Janus particle synthesis and assembly. *Adv. Mater*. **22,** 1060–1071 (2010).
21. Glotzer, S. C. & Solomon, M. J. Anisotropy of building blocks and their assembly into complex structures. *Nat. Mater*. **6,** 557–562 (2007).
22. Dommersnes, P. *et al.* Active structuring of colloidal armour on liquid drops. *Nat. Commun*. **4,** 2066 (2013).
23. Rozynek, Z., Mikkelsen, A., Dommersnes, P. & Fossum, J. O. Electroformation of Janus and patchy capsules. *Nat. Commun*. **5,** 3945 (2014).
24. Pontani, L. L., Haase, M. F., Raczkowska, I. & Brujic, J. Immiscible lipids control the morphology of patchy emulsions. *Soft Matter* **9,** 7150–7157 (2013).
25. Gross, A. F., Sherman, E. & Vajo, J. J. Aqueous room temperature synthesis of cobalt and zinc sodalite zeolitic imidizolate frameworks. *Dalton Trans*. **41,** 5458–5460 (2012).
26. Lach, S., Yoon, S. M. & Grzybowski, B. A. Tactic, reactive and functional droplets outside of equilibrium. *Chem. Soc. Rev*. **45,** 4766–4796 (2016).
27. Zarzar, L. D. *et al.* Dynamically reconfigurable complex emulsions via tunable interfacial tensions. *Nature* **518,** 520–524 (2015).
28. Davies Wykes, M. S. *et al.* Dynamic self-assembly of microscale rotors and swimmers. *Soft Matter* **12,** 4584–4589 (2016).
29. Paven, M. *et al.* Light-driven delivery and release of materials using liquid marbles. *Adv. Funct. Mater*. **26,** 3199–3206 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge support from the Institute for Basic Science Korea, project code IBS-R020-D1.

Author Contributions Z.Y. and J.W. performed most of the experiments. Y.I.S. performed some experiments and developed the theoretical models. B.A.G. conceived the project and supervised the research. All authors participated in writing the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to B.A.G. (nanogrzybowski@gmail.com).

Reviewer Information *Nature* thanks W. Kunz and the other anonymous reviewer(s) for their contribution to the peer review of this work.

doi:10.1038/nature25450

Emergent constraint on equilibrium climate sensitivity from global temperature variability

Peter M. Cox[1], Chris Huntingford[2] & Mark S. Williamson[1]

Equilibrium climate sensitivity (ECS) remains one of the most important unknowns in climate change science. ECS is defined as the global mean warming that would occur if the atmospheric carbon dioxide (CO_2) concentration were instantly doubled and the climate were then brought to equilibrium with that new level of CO_2. Despite its rather idealized definition, ECS has continuing relevance for international climate change agreements, which are often framed in terms of stabilization of global warming relative to the pre-industrial climate. However, the 'likely' range of ECS as stated by the Intergovernmental Panel on Climate Change (IPCC) has remained at 1.5–4.5 degrees Celsius for more than 25 years[1]. The possibility of a value of ECS towards the upper end of this range reduces the feasibility of avoiding 2 degrees Celsius of global warming, as required by the Paris Agreement. Here we present a new emergent constraint on ECS that yields a central estimate of 2.8 degrees Celsius with 66 per cent confidence limits (equivalent to the IPCC 'likely' range) of 2.2–3.4 degrees Celsius. Our approach is to focus on the variability of temperature about long-term historical warming, rather than on the warming trend itself. We use an ensemble of climate models to define an emergent relationship[2] between ECS and a theoretically informed metric of global temperature variability. This metric of variability can also be calculated from observational records of global warming[3], which enables tighter constraints to be placed on ECS, reducing the probability of ECS being less than 1.5 degrees Celsius to less than 3 per cent, and the probability of ECS exceeding 4.5 degrees Celsius to less than 1 per cent.

Many attempts have been made to constrain ECS, typically using either the record of historical warming or reconstructions of past climates[4]. Methods based on historical warming are affected by uncertainties in ocean heat uptake and the contribution of aerosols to net radiative forcing[5,6]. These methods also diagnose the effective climate sensitivity over the historical period, which may be different to ECS, owing to the strength of climate feedbacks varying with the evolving pattern of surface temperature change[4,7–9]. Although methods based on past climatic periods, such as the Last Glacial Maximum[10], are more closely related to the concept of equilibrium, they suffer instead from even larger uncertainties in the reconstruction of net radiative forcing.

As an alternative, the emergent constraint approach uses an ensemble of complex Earth system models to estimate the relationship between a modelled but observable variation in the Earth system and a predicted future change[2,11]. The model-derived emergent relationship can then be combined with the quantification of the observed variation to produce an emergent constraint on the predicted future change[2,11,12]. Here we present an emergent constraint on ECS that is based on the variability of global-mean temperature.

To inform our search for an emergent constraint, we consider the simple 'Hasselmann model'[13] for the variation in global mean temperature ΔT in response to a radiative forcing Q:

$$C\frac{\mathrm{d}\Delta T}{\mathrm{d}t} = Q - \lambda\Delta T = N \qquad (1)$$

The constant heat capacity C in this model is a simplification that is known to be a poor representation of ocean heat uptake on longer timescales[14–16]. However, we find that it still offers very useful guidance about global temperature variability on shorter timescales. The climate

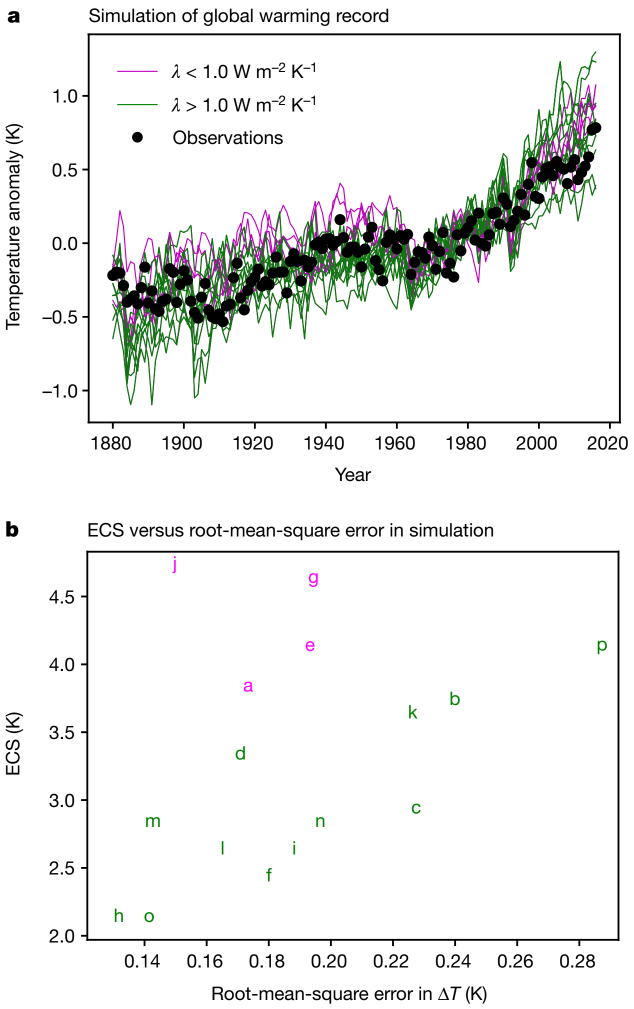


Figure 1 | Historical global warming. a, Simulated change in global temperature from 16 CMIP5 models (coloured lines), compared to the global temperature anomaly from the HadCRUT4 dataset (black dots). The anomalies are relative to a baseline period of 1961–1990. The model lines are colour-coded, with lower-sensitivity models ($\lambda > 1$ W m^{-2} K^{-1}) shown by green lines and higher-sensitivity models ($\lambda < 1$ W m^{-2} K^{-1}) shown by magenta lines. **b**, Scatter plot of each model's ECS against the root-mean-square error in the fit of each model to the observational record. Individual CMIP5 model runs are denoted by the letters listed in Extended Data Table 1.

[1]College of Engineering, Mathematics and Physical Science, University of Exeter, Exeter EX4 4QF, UK. [2]Centre for Ecology and Hydrology, Wallingford OX10 8BB, UK.

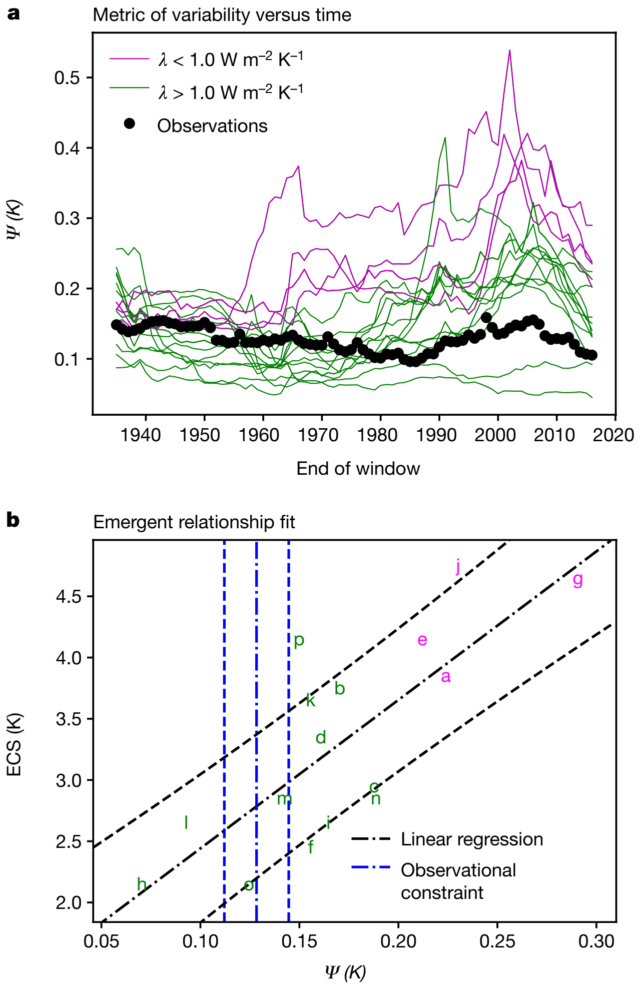


Figure 2 | Metric of global mean temperature variability. a, Ψ metric of variability versus time, from the CMIP5 models (coloured lines), and the HadCRUT4 observational data (black circles). The Ψ values are calculated for windows of width 55 yr, after linear de-trending in each window. These 55-yr windows are shown for different end times. As in Fig. 1, lower-sensitivity models ($\lambda > 1$ W m^{-2} K^{-1}) are shown by green lines and higher-sensitivity models ($\lambda < 1$ W m^{-2} K^{-1}) are shown by magenta lines. **b**, Emergent relationship between ECS and the Ψ metric. The black dot-dashed line shows the best-fit linear regression across the model ensemble, with the prediction error for the fit given by the black dashed lines (see Methods). The vertical blue lines show the observational constraint from the HadCRUT4 observations: the mean (dot-dashed line) and the mean plus and minus one standard deviation (dashed lines).

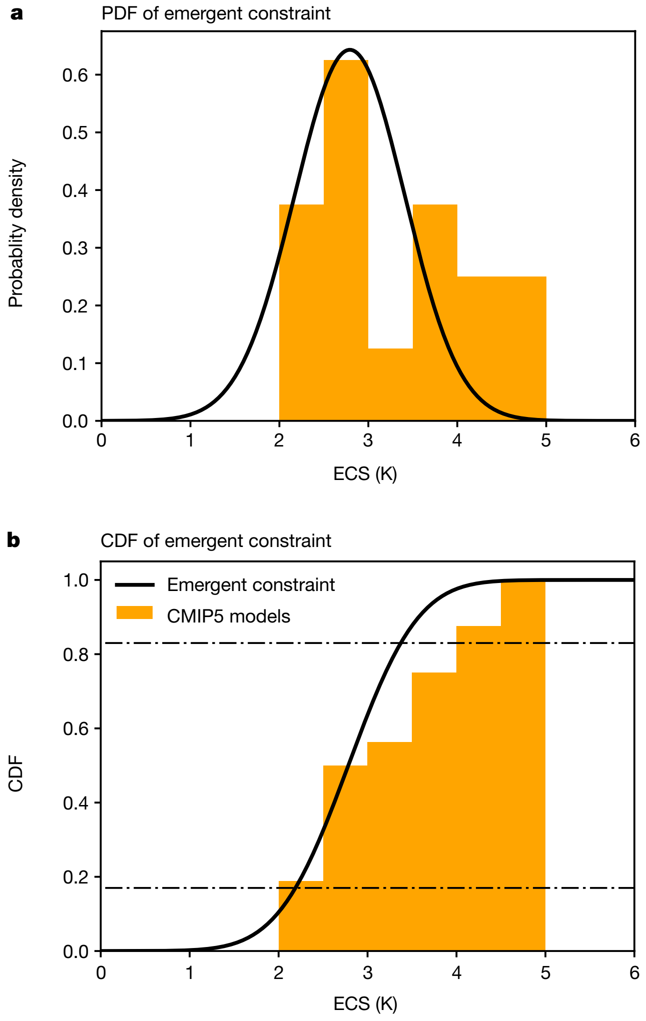


Figure 3 | Emergent constraint on ECS. a, The PDF for ECS. **b**, The related CDF. The horizontal dot-dashed lines show the 66% confidence limits on the CDF plot. The orange histograms (both panels) show the prior distributions that arise from equal weighting of the CMIP5 models in 0.5 K bins.

feedback factor λ determines how the net top-of-atmosphere planetary energy balance N varies with temperature change ΔT in response to a radiative forcing change Q. ECS and λ are inversely related, with a constant of proportionality that is the radiative forcing due to doubling of atmospheric CO$_2$, $Q_{2\times\mathrm{CO2}}$ so that ECS $= Q_{2\times\mathrm{CO2}}/\lambda$. Although the diagnosed $Q_{2\times\mathrm{CO2}}$ varies across the model ensemble[17], the uncertainty in ECS is predominantly due to uncertainty in λ, which varies from 0.6 W m^{-2} K^{-1} to 1.8 W m^{-2} K^{-1}, as shown in Extended Data Table 1.

If Q can be approximated as white-noise forcing with variance σ_Q^2, the Hasselmann model can be solved to give expressions for the variance of global temperature σ_T^2 and the one-year-lag autocorrelation of the global temperature α_{1T}, which can be combined to yield an equation for ECS (see Methods):

$$\mathrm{ECS} = \sqrt{2}\,Q_{2\times\mathrm{CO2}}\left\{\frac{\sigma_T}{\sigma_Q}\right\}\frac{1}{\sqrt{-\log_e \alpha_{1T}}} = \sqrt{2}\,\frac{Q_{2\times\mathrm{CO2}}}{\sigma_Q}\,\Psi \qquad (2)$$

where $\Psi = \sigma_T/\sqrt{-\log_e \alpha_{1T}}$ is our key metric of global temperature variability. This equation is essentially a fluctuation–dissipation relationship[18] relating the variability of the climate (σ_Q, σ_T, α_{1T}) to its sensitivity to external forcing (ECS).

Observational records of global mean temperature change[3] enable Ψ to be estimated for the real world. The variance of the net radiative forcing is approximately equal to the variance of the top-of-the-atmosphere flux σ_N^2, which can in principle be estimated from satellite measurements. However, the available satellite records are currently too short to provide reliable estimates of σ_N. In addition, the radiative forcing due to doubling CO$_2$ ($Q_{2\times\mathrm{CO2}}$) is not observable in the real world. This means that the right-hand side of equation (2) cannot be directly estimated from observations. Fortunately, we find that the variation in ECS is weakly correlated with $Q_{2\times\mathrm{CO2}}/\sigma_N$ across the model ensemble (see Extended Data Table 1). We can therefore approximate the predicted gradient of the ECS versus Ψ emergent relationship using the ensemble mean value of $Q_{2\times\mathrm{CO2}}/\sigma_N$ ($= 8.7$). Our theory therefore predicts a gradient of the ECS versus Ψ emergent relationship of $8.7\sqrt{2} = 12.2$.

Figure 1a shows the simulation of global warming in the historical simulations with the 16 models in the CMIP5 ensemble[19,20] used here (see list in Extended Data Table 1). Here and throughout, higher-sensitivity models ($\lambda < 1.0$ W m^{-2} K^{-1}) are shown in magenta and lower-sensitivity models ($\lambda > 1.0\,\mathrm{W\,m^{-2}\,K^{-1}}$) are shown in green. Observations from the HadCRUT4 dataset[3] are shown by the black line marked with dots. Figure 1a illustrates that both high- and low-sensitivity models are able to fit the historical record with reasonable fidelity, despite implying very different future climates. Models with higher ECS values also have longer response times, and there are variations across the models in net radiative forcing and in ocean heat uptake—allowing models with both high and low sensitivities to reproduce historical global warming[21]. As a result, the fit to the global temperature record does not provide a direct constraint on ECS, as shown in Fig. 1b.

To test whether variability is a better constraint on ECS, we de-trend the global mean temperature records from the models and the observations. Our approach to de-trending is informed by techniques designed to detect precursors of potential tipping points[22] such as 'critical slowing down'[23]. The method applied in that case is to use a moving window, to linearly de-trend within that window, and then to calculate statistics of the de-trended residuals. For tipping point detection, the favoured variable is often the autocorrelation, which measures the memory in fluctuations of the analysed variable[23]. We use a similar approach, although here we apply it to analyse the relationship between Ψ and ECS across the ensemble of models, rather than to detect declining system resilience in a single realization of the system.

We analyse the annual-mean global-mean temperature time series from 16 CMIP5 historical simulations and compare to the HadCRUT4 observational dataset. Although there were another 23 historical runs available in the CMIP5 archive, we chose to use just one model variant from each climate centre, to avoid biasing the emergent constraint towards the centres with the most model runs in the archive. Where there was more than one model variant from a modelling centre, we took the model variant from that centre that had the smallest root-mean-square (r.m.s.) error in the fit to the record of observed global warming from 1880 to 2016. The remaining 23 model runs (which included some initial condition ensembles) were subsequently used to test the robustness of the emergent constraint (see Extended Data Fig. 1).

Figure 2a shows the resulting variation in Ψ for each of the models and the observations, using a window width of 55 yr, and data from 1880 to 2016 to match the available observational datasets. Although Ψ varies in time, the different models are clearly distinguished, in contrast to the simulations of historical global warming (Fig. 1a). In particular, the Ψ values separate higher-sensitivity models (magenta lines) from lower-sensitivity models (green lines), with higher-sensitivity models producing larger Ψ values. It is also worth noting that Ψ from the observational data are within the range of the lower-sensitivity models but clearly outside the range of the higher-sensitivity models. Figure 2b shows the emergent relationship between ECS and the time-mean Ψ values across the model ensemble, with a best-fit gradient that is very close to our theoretical value. The vertical blue lines show the observational constraint on Ψ from the HadCRUT4 dataset, but similar observational constraints are also derived from other datasets of global mean temperature (see Extended Data Table 2).

As in previous studies[11,12] the emergent relationship from the historical runs and observational constraint can be combined to provide an emergent constraint on ECS. This involves convolving the prediction error implied by the fit of the scatter plot to the emergent relationship, with the uncertainty in the observations, to produce a probability density function (PDF) for the y-axis variable (see Methods). Figure 3a shows the resulting PDF for ECS (black curve). For comparison, the prior PDF implied by the equal-weighted model ensemble is shown by the orange histogram. The emergent constraint PDF is sharply peaked around a best estimate of ECS = 2.8 K, which is slightly smaller than the centre of the IPCC range of 1.5–4.5 K. Our best estimate of ECS is considerably larger than the values derived from raw energy budget constraints[8,24,25] but similar to some recent estimates that account for time-dependent and forcing-dependent feedbacks[9,26].

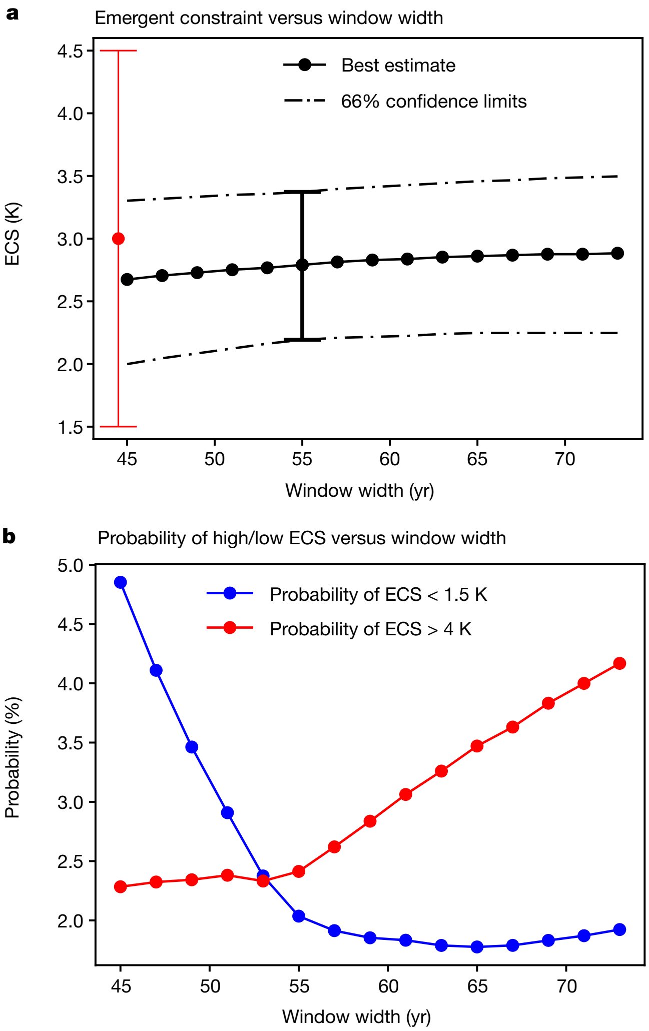


Figure 4 | Sensitivity of the emergent constraint on ECS to window width. a, Central estimate and 66% confidence limits. The thick black bar shows the minimum uncertainty at a window width of 55 yr and the red bar shows the equivalent 'likely' IPCC range of 1.5–4.5 K. **b**, Probabilities of ECS > 4 K (red line and symbols) and ECS < 1.5 K (blue line and symbols).

Figure 3b shows the resulting cumulative density function (CDF), which gives the probability of ECS taking a value lower than the value shown on the x axis. The black horizontal lines in Fig. 3b show the 66% confidence limits (2.2 K to 3.4 K), or approximately 2.8 ± 0.6 K. Relative to the IPCC range of 1.5–4.5 K, this constraint on ECS therefore reduces the uncertainty by about 60%. Indeed, even the 95% confidence limits from the emergent constraint (1.6 K to 4.0 K) fit well within the IPCC 'likely' range for ECS. Our constraint is therefore at odds with a suggestion that the lower 66% confidence limit for ECS could be as high as 3 K (ref. 27). If we instead use all 39 historical runs in the CMIP5 archive, we find a slightly weaker emergent relationship, but derive a very similar emergent constraint on ECS (Extended Data Table 2). The constraint is also robust to the choice of observational dataset, and to whether or not the model global temperature is calculated just across the points where there were observations[28] (Extended Data Table 2 and Extended Data Fig. 2).

Our choice of window width was informed by sensitivity studies in which the emergent constraint was calculated for a range of this parameter. Figure 4a shows the best estimate and 66% confidence limits on ECS as a function of the width of the de-trending window. Our best estimate is relatively insensitive to the chosen window width, but the 66% confidence limits show a greater sensitivity, with the minimum in uncertainty at a window width of about 55 yr (as used in the analysis above). As Extended Data Fig. 3 shows, at this optimum window width the best-fit gradient of the emergent relationship between ECS and Ψ (=12.1) is also very close to our theory-predicted value of $\sqrt{2}\,Q_{2\times\mathrm{CO2}}/\sigma_Q$ (=12.2). This might be expected if this window length optimally separates forced trend from variability.

Figure 4b shows the probability of ECS > 4 K and ECS < 1.5 K as a function of window width. For comparison, the IPCC 'likely' range of 1.5–4.5 K implies a 25% probability of ECS > 4 K, and a 16% probability of ECS < 1.5 K. At the optimum window width of 55 yr, both probabilities are close to their minimum values of less than 2.5%. Our emergent constraint therefore greatly reduces the uncertainty in the ECS value of Earth's climate, implying a less than 1 in 40 chance of ECS > 4 K, and renewing hope that we may yet be able to avoid global warming exceeding 2 K.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 July; accepted 13 December 2017.

1. Collins, M. *et al.* Long-term climate change: projections, commitments and irreversibility. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) Ch. 12 (Cambridge Univ. Press, 2013).
2. Hall, A. & Qu, X. Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett.* **33,** L03502 (2006).
3. Morice, C. P. *et al.* Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 dataset. *J. Geophys. Res.* **117,** D08101 (2012).
4. Knutti, R. *et al.* Beyond equilibrium climate sensitivity. *Nat. Geosci.* **10,** 727–736 (2017).
5. Gregory, J. M. *et al.* An observationally based estimate of the climate sensitivity. *J. Clim.* **15,** 3117–3121 (2002).
6. Forster, P. M. *et al.* Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models. *J. Geophys. Res. Atmos.* **118,** (2013).
7. Gregory, J. M. & Andrews, T. Variation in climate sensitivity and feedback parameters during the historical period. *Geophys. Res. Lett.* **43,** 3911–3920 (2016).
8. Forster, P. M. Inference of climate sensitivity from analysis of Earth's radiation budget. *Ann. Rev. Earth Planet. Sci*. **44,** 85–106 (2016).
9. Armour, K. C. Energy budget constraints on climate sensitivity in light of inconstant climate feedbacks. *Nat. Clim. Chang*. **7,** 331–335 (2017).
10. Annan, J. D. & Hargreaves, J. C. Using multiple observationally-based constraints to estimate climate sensitivity. *Geophys. Res. Lett.* **33,** L06704 (2006).
11. Cox, P. M. *et al.* Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature* **494,** 341–344 (2013).
12. Wenzel, S. *et al.* Projected land photosynthesis constrained by changes in the seasonal cycle of atmospheric CO_2. *Nature* **538,** 499–501 (2016).
13. Hasselmann, K. Stochastic climate models. I. Theory. *Tellus* **28,** 473–485 (1976).
14. MacMynowski, D. G. *et al.* The frequency response of temperature and precipitation in a climate model. *Geophys. Res. Lett.* **38,** L16711 (2011).
15. Caldeira, K. & Myhrvold, N. P. Projections of the pace of warming following an abrupt increase in atmospheric carbon dioxide concentration. *Environ. Res. Lett.* **8,** 034039 (2013).
16. Geoffroy, O. *et al.* Transient climate response in a two-layer energy-balance model. Part I: Analytical solution and parameter calibration using CMIP5 AOGCM experiments. *J. Clim*. **26,** 1841–1857 (2013).
17. Flato, G. *et al.* Evaluation of climate models. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) Ch. 9 (Cambridge Univ. Press, 2013).
18. Leith, C. E. Climate response and fluctuation dissipation. *J. Atmos. Sci*. **32,** 2022–2026 (1975).
19. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc*. **93,** 485–498 (2012).
20. Andrews, T. *et al.* Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean models. *Geophys. Res. Lett.* **39,** L09712 (2012).
21. Kiehl, J. T. Twentieth century climate model response and climate sensitivity. *Geophys. Res. Lett.* **34,** L22710 (2007).
22. Lenton, T. M. *et al.* Tipping elements in the Earth's climate system. *Proc. Natl Acad. Sci. USA* **105,** 1786–1793 (2008).
23. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461,** 53–59 (2009).
24. Otto, A. *et al.* Energy budget constraints on climate response. *Nat. Geosci.* **6,** 415–416 (2013).
25. Lewis, N. & Curry, J. A. The implications for climate sensitivity of AR5 forcing and heat uptake estimates. *Clim. Dyn*. **45,** 1009–1023 (2015).
26. Marvel, K. *et al.* Implications for climate sensitivity from the response to individual forcings. *Nat. Clim. Chang*. **6,** 386–389 (2015).
27. Sherwood, S. C., Bony, S. & Dufresne, J.-L. Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature* **505,** 37–42 (2014).
28. Cowtan, K. & Way, R. G. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Q. J. R. Meteorol. Soc*. **140,** 1935–1944 (2014).

Acknowledgements This work was supported by the European Research Council (ERC) ECCLES project, grant agreement number 742472 (P.M.C.); the EU Horizon 2020 Research Programme CRESCENDO project, grant agreement number 641816 (P.M.C. and M.S.W.); the EPSRC-funded ReCoVER project (M.S.W.); and the NERC CEH National Capability fund (C.H.). We also acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups (listed in Extended Data Table 1 of this paper) for producing and making available their model output.

Author Contributions All authors collaboratively designed the study and contributed to the manuscript. P.M.C. led the study and drafted the manuscript. C.H. was the lead on the time-series data for the CMIP5 models. M.S.W. led on the theoretical analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to P.M.C. (p.m.cox@exeter.ac.uk).

Reviewer Information *Nature* thanks P. Forster and T. Mauritsen for their contribution to the peer review of this work.

METHODS

Theoretical basis for the emergent relationship. We hypothesize that equation (1) (the 'Hasselmann model') is a reasonable approximation to the short-term variability of the global mean temperature anomaly ΔT:

$$C\frac{\mathrm{d}\Delta T}{\mathrm{d}t}+\lambda\Delta T=Q \tag{3}$$

If trends arising from net radiative forcing and ocean heat uptake can be successfully removed, the net radiative forcing term Q can be approximated by white noise. Under these circumstances, equation (1) is essentially the Ornstein–Uhlenbeck equation, which describes Brownian motion, and has standard solutions (for example, see https://en.wikipedia.org/wiki/Ornstein–Uhlenbeck_process) for the lag-one-year autocorrelation of the temperature:

$$\alpha_{1T}=\exp\left\{-\frac{\lambda}{C}\right\} \tag{4}$$

and the ratio of the variances of T and Q:

$$\frac{\sigma_T^2}{\sigma_Q^2}=\frac{1}{2\lambda C} \tag{5}$$

These two equations can be combined to eliminate the unknown heat capacity C and therefore to provide an expression for the climate feedback factor λ:

$$\lambda=\left\{\frac{\sigma_Q}{\sigma_T}\right\}\sqrt{-\frac{1}{2}\log_{\mathrm{e}}\alpha_{1T}} \tag{6}$$

The ECS and λ are inversely related by a constant of proportionality, which is the radiative forcing due to doubling of atmospheric CO_2 ($Q_{2\times\mathrm{CO2}}$), so that $\mathrm{ECS}=Q_{2\times\mathrm{CO2}}/\lambda$. Thus, we can also derive an expression for ECS in terms of the variability of T and Q:

$$\mathrm{ECS}=Q_{2\times\mathrm{CO2}}\left\{\frac{\sigma_T}{\sigma_Q}\right\}\sqrt{\frac{2}{-\log_{\mathrm{e}}\alpha_{1T}}} \tag{7}$$

Least-squares linear regression. Least-squares linear regressions were calculated using well established formulae (see for example http://mathworld.wolfram.com/LeastSquaresFitting.html). The linear regression f_n between a time series given by y_n and a time series given by x_n is defined by a gradient b and intercept a:

$$f_n=a+bx_n \tag{8}$$

Minimizing the least-squares error for y_n involves minimizing:

$$s^2=\frac{1}{N-2}\sum_{n=1}^{N}\{y_n-f_n\}^2 \tag{9}$$

where N is the number of data points in each time series. In this case, the best-fit gradient is given by:

$$\bar{b}=\frac{\sigma_{xy}^2}{\sigma_x^2} \tag{10}$$

Here $\sigma_x^2=\sum_{n=1}^{N}\{x_n-\bar{x}\}^2/N$ is the variance of x_n and $\sigma_{xy}^2=\sum_{n=1}^{N}\{x_n-\bar{x}\}\times\{y_n-\bar{y}\}/N$ is the covariance of the x_n and y_n time series, with means of $\bar{x}$ and $\bar{y}$, respectively. The standard error of b is given by:

$$\sigma_b=\frac{s}{\sigma_x\sqrt{N}} \tag{11}$$

which defines a Gaussian probability density for b:

$$P(b)=\frac{1}{\sqrt{2\pi\sigma_b^2}}\exp\left\{-\frac{(b-\bar{b})^2}{2\sigma_b^2}\right\} \tag{12}$$

Finally, the 'prediction error' of the regression is the following function of x:

$$\sigma_f(x)=s\sqrt{1+\frac{1}{N}+\frac{\{x-\bar{x}\}^2}{N\sigma_x^2}} \tag{13}$$

This expression defines contours of equal probability density around the best-fit linear regression, which represent the probability density of y given x:

$$P\{y|x\}=\frac{1}{\sqrt{2\pi\sigma_f^2}}\exp\left\{-\frac{(y-f(x))^2}{2\sigma_f^2}\right\} \tag{14}$$

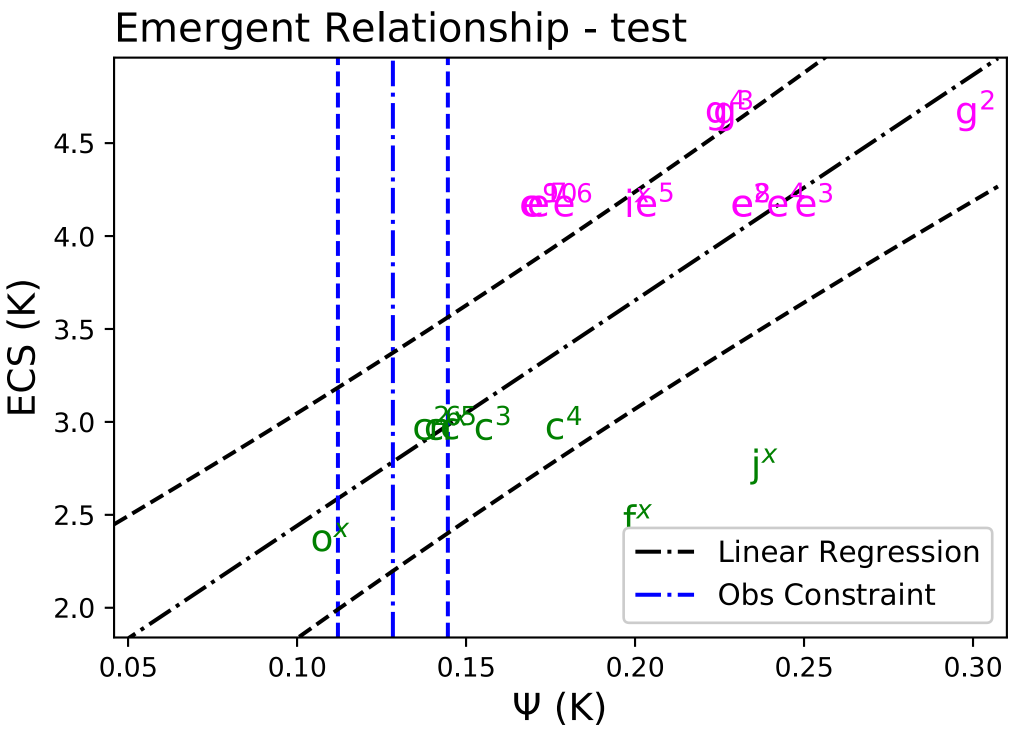
where $\sigma_f=\sigma_f(x)$, as described above.

Calculation of the PDF for ECS. The emergent constraint derived in this study is a linear regression across the CMIP5 models between ECS and the Ψ statistic of the de-trended global temperature. In the context of the least-squares linear regression presented above, ECS is equivalent to y, and Ψ is equivalent to x. The linear regression therefore provides an equation for the probability of ECS given Ψ (that is, the equation for $P\{y|x\}$ above). In addition, the Ψ statistic calculated from the de-trended observational dataset provides an observation-based PDF for Ψ. Given these two PDFs, $P\{\mathrm{ECS}|\Psi\}$ and $P(\Psi)$, the PDF for ECS is calculated by numerically integrating:

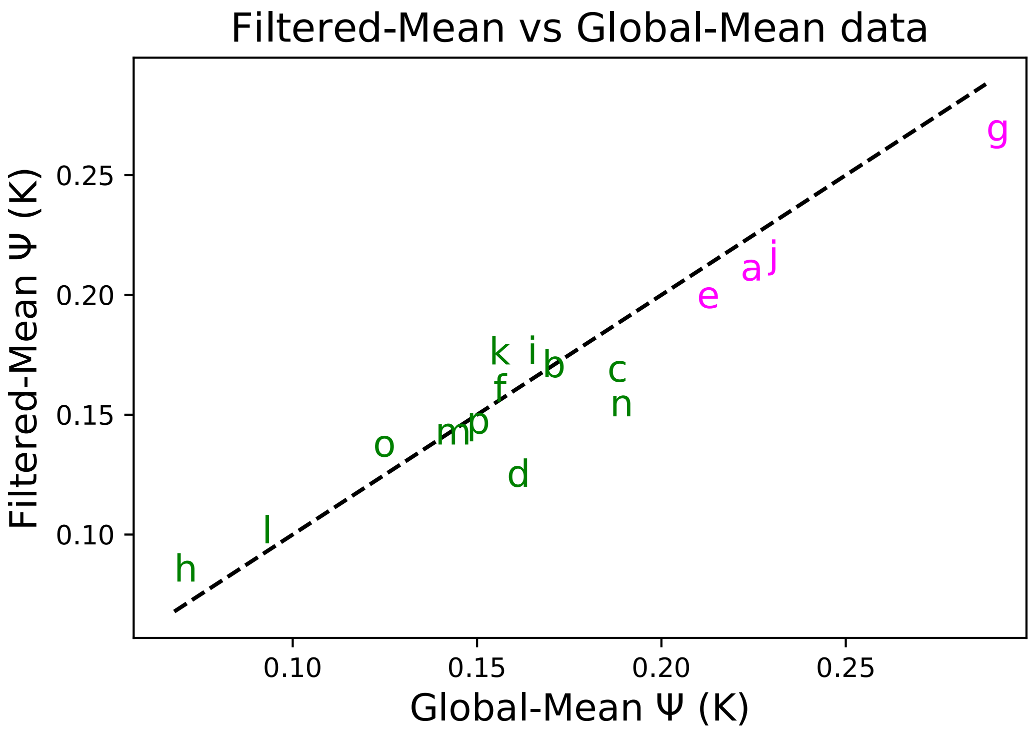
$$P(\mathrm{ECS})=\int_{-\infty}^{\infty}P\{\mathrm{ECS}|\Psi\}\,P(\Psi)\mathrm{d}\Psi \tag{15}$$

Data availability. The datasets generated during the current study are available from the corresponding author on reasonable request.

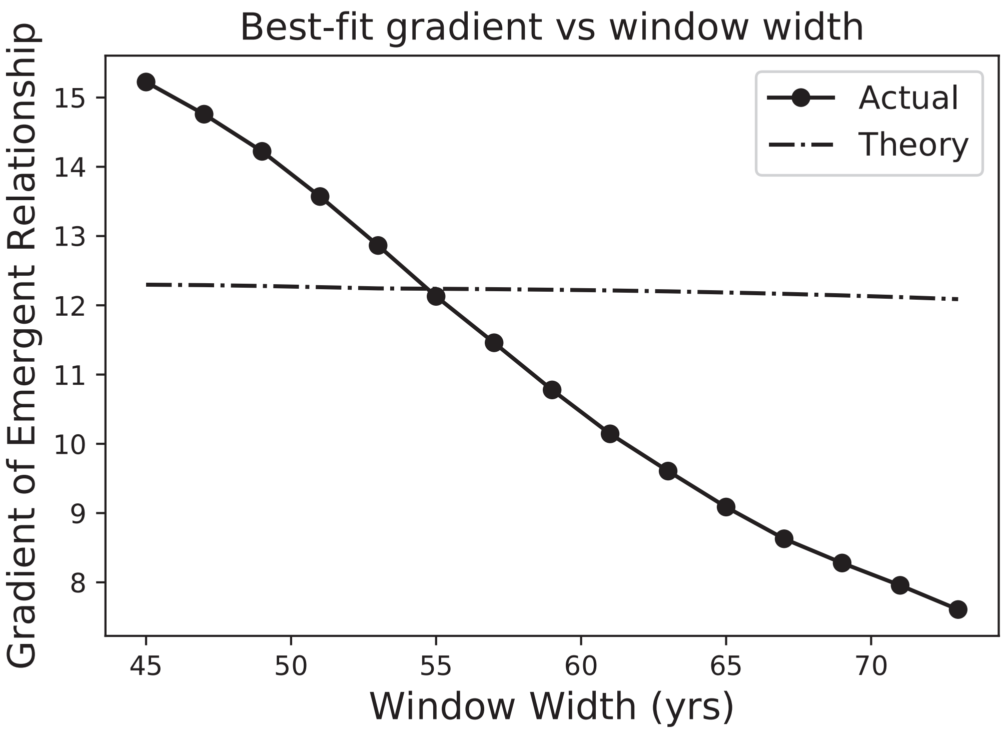
Code availability. The Python code used to produce the figures in this paper is available from the corresponding author on reasonable request.



Extended Data Figure 1 | Test of emergent relationship against models not used in the calibration. The test set includes additional models from some climate centres (labelled ‘f^x’, ‘f^y’ and so on), and initial condition ensembles with particular models (labelled ‘c^2’, ‘c^3’ and so on). The black dot-dashed line shows the best-fit linear regression across the model ensemble, with the prediction error for the fit given by the black dashed lines (see Methods). The vertical blue lines show the observational constraint from the HadCRUT4 observations: the mean (dot-dashed line) and the mean plus and minus one standard deviation (dashed lines). Individual CMIP5 model runs are denoted by the letters listed in Extended Data Table 1.

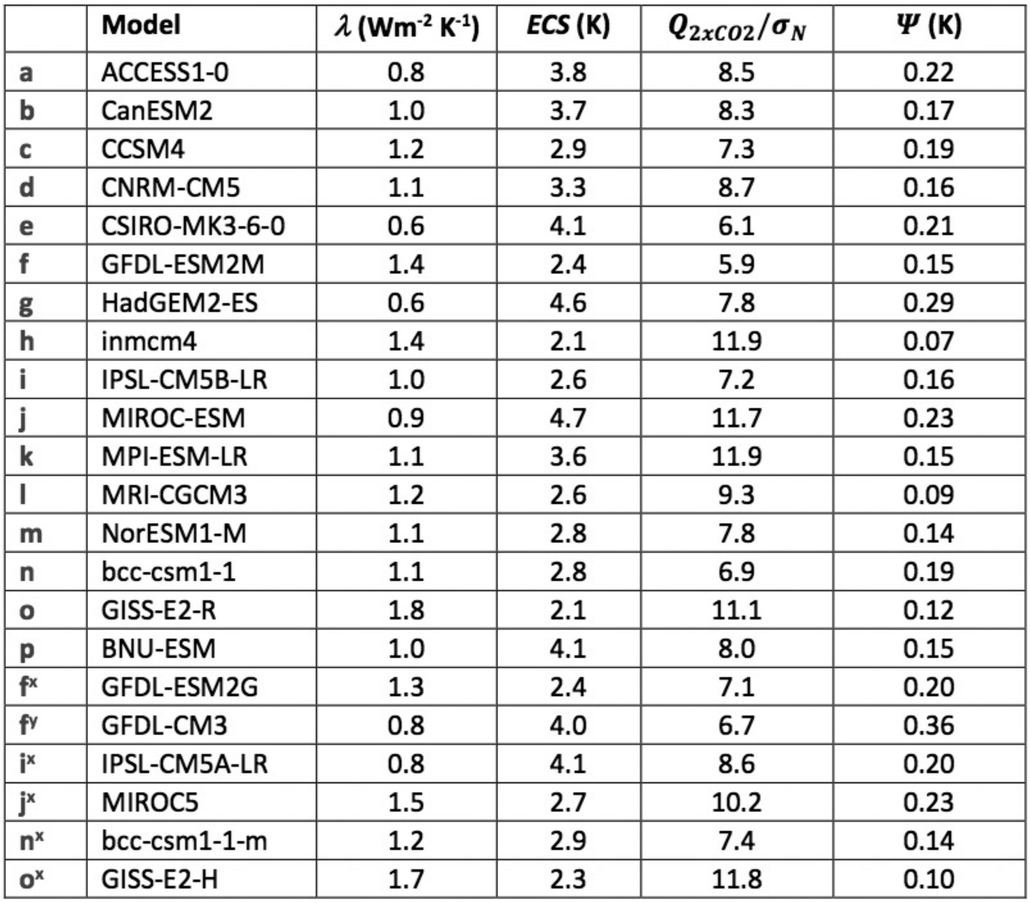


Extended Data Figure 2 | Comparison of Ψ statistics for the 16 CMIP5 models from 'filtered-mean' temperature and global-mean temperature. The filtered model output calculates area-mean values of temperature using only the points where there are observations in the HadCRUT4 dataset. All cases analyse 1880–2016 and use a 55-yr window width. The dotted line is the 1:1 line.



Extended Data Figure 3 | Gradient of emergent relationship between ECS and Ψ as a function of window width. The dotted line shows the gradient predicted with equation (2) using the ensemble-mean value of $Q_{2\times\mathrm{CO2}}/\sigma_N$. Note that the theory (dot-dashed line) fits best at the optimal window width of 55 yr. All cases here analyse 1880–2016 and use the 16-model ensemble.

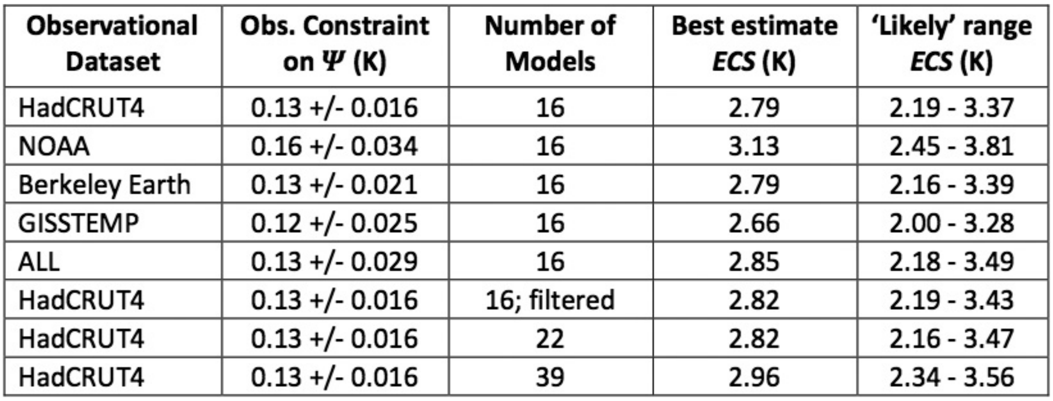
Extended Data Table 1 | Earth system models used in this study, as provided by the CMIP5 project[19]



	Model	λ (Wm^{-2} K^{-1})	*ECS* (K)	Q_{2xCO2}/σ_N	Ψ (K)
a	ACCESS1-0	0.8	3.8	8.5	0.22
b	CanESM2	1.0	3.7	8.3	0.17
c	CCSM4	1.2	2.9	7.3	0.19
d	CNRM-CM5	1.1	3.3	8.7	0.16
e	CSIRO-MK3-6-0	0.6	4.1	6.1	0.21
f	GFDL-ESM2M	1.4	2.4	5.9	0.15
g	HadGEM2-ES	0.6	4.6	7.8	0.29
h	inmcm4	1.4	2.1	11.9	0.07
i	IPSL-CM5B-LR	1.0	2.6	7.2	0.16
j	MIROC-ESM	0.9	4.7	11.7	0.23
k	MPI-ESM-LR	1.1	3.6	11.9	0.15
l	MRI-CGCM3	1.2	2.6	9.3	0.09
m	NorESM1-M	1.1	2.8	7.8	0.14
n	bcc-csm1-1	1.1	2.8	6.9	0.19
o	GISS-E2-R	1.8	2.1	11.1	0.12
p	BNU-ESM	1.0	4.1	8.0	0.15
f^x	GFDL-ESM2G	1.3	2.4	7.1	0.20
f^y	GFDL-CM3	0.8	4.0	6.7	0.36
i^x	IPSL-CM5A-LR	0.8	4.1	8.6	0.20
j^x	MIROC5	1.5	2.7	10.2	0.23
n^x	bcc-csm1-1-m	1.2	2.9	7.4	0.14
o^x	GISS-E2-H	1.7	2.3	11.8	0.10

The first column shows the symbol used for each model in Figs 1b and 2b. The third and fourth columns list λ and the ECS values as given in IPCC AR5 table 9.5 (ref. 17). The fifth and sixth columns show statistics calculated in this study for the period 1880–2016 and using a window width of 55 yr. The fifth column shows the ratio of the radiative forcing due to doubling CO_2 ($Q_{2\times CO2}$) to the standard deviation of the net top-of-atmosphere flux σ_N; and the sixth column shows the time-mean Ψ statistic for each model.

Extended Data Table 2 | Robustness of the emergent constraint to the choice of observational dataset and model ensemble



Observational Dataset	Obs. Constraint on Ψ (K)	Number of Models	Best estimate *ECS* (K)	'Likely' range *ECS* (K)
HadCRUT4	0.13 +/- 0.016	16	2.79	2.19 - 3.37
NOAA	0.16 +/- 0.034	16	3.13	2.45 - 3.81
Berkeley Earth	0.13 +/- 0.021	16	2.79	2.16 - 3.39
GISSTEMP	0.12 +/- 0.025	16	2.66	2.00 - 3.28
ALL	0.13 +/- 0.029	16	2.85	2.18 - 3.49
HadCRUT4	0.13 +/- 0.016	16; filtered	2.82	2.19 - 3.43
HadCRUT4	0.13 +/- 0.016	22	2.82	2.16 - 3.47
HadCRUT4	0.13 +/- 0.016	39	2.96	2.34 - 3.56

The 'ALL' dataset takes the mean and standard deviation of the Ψ values for all four global-mean temperature datasets (by concatenating the individual Ψ time series). The 'filtered' model output calculates area-mean values of temperature just using the points where there are observations in the HadCRUT4 dataset[27]. All cases analyse 1880–2016 and use a 55-yr window width.

doi:10.1038/nature25009

A record of deep-ocean dissolved O_2 from the oxidation state of iron in submarine basalts

Daniel A. Stolper[1] & C. Brenhin Keller[1,2]

The oxygenation of the deep ocean in the geological past has been associated with a rise in the partial pressure of atmospheric molecular oxygen (O_2) to near-present levels and the emergence of modern marine biogeochemical cycles[1–5]. It has also been linked to the origination and diversification of early animals[3,5–7]. It is generally thought that the deep ocean was largely anoxic from about 2,500 to 800 million years ago[1–12], with estimates of the occurrence of deep-ocean oxygenation and the linked increase in the partial pressure of atmospheric oxygen to levels sufficient for this oxygenation ranging from about 800 to 400 million years ago[3,5,7,11,13]. Deep-ocean dissolved oxygen concentrations over this interval are typically estimated using geochemical signatures preserved in ancient continental shelf or slope sediments, which only indirectly reflect the geochemical state of the deep ocean. Here we present a record that more directly reflects deep-ocean oxygen concentrations, based on the ratio of Fe^{3+} to total Fe in hydrothermally altered basalts formed in ocean basins. Our data allow for quantitative estimates of deep-ocean dissolved oxygen concentrations from 3.5 billion years ago to 14 million years ago and suggest that deep-ocean oxygenation occurred in the Phanerozoic (541 million years ago to the present) and potentially not until the late Palaeozoic (less than 420 million years ago).

There is general agreement that between about 2,500 and 2,300 million years (Myr) ago, the partial pressure of oxygen in the atmosphere, $p_{O_2,atm}$, rose from below to above approximately 10^{-5} times present atmospheric levels (PAL)[1,2,4], and since about 400 Myr ago, $p_{O_2,atm}$ has remained above about 70% PAL[10,14,15]. Biogeochemical models suggest that $p_{O_2,atm}$ levels exceeding 15% (ref. 3) to 50% (ref. 1) PAL are needed to oxygenate the deep ocean. From about 2,500–800 Myr ago, it is generally thought[1–12] that $p_{O_2,atm}$ was below this threshold and that the deep ocean was either anoxic or contained at most a few micromoles of O_2 molecules per kilogram of seawater (μmol kg^{-1}) (modern deep-ocean O_2 concentrations average about 180 μmol kg^{-1})[16]. Estimates of when $p_{O_2,atm}$ exceeded these levels and the deep ocean became oxygenated range from 815 to 400 Myr ago[3,5,7,11,13]. Constraining the timing of deep-ocean oxygenation is important because it signals the beginning of modern marine biogeochemical cycles and has been causally linked (owing to the implied rise in $p_{O_2,atm}$) to the Neoproterozoic origination of animals (which require O_2)[3,5–7,9,10].

Most reconstructions of Mesozoic and older deep-ocean geochemical conditions are based on the chemical and isotopic composition of continental shelf and slope sediments[3–5,9–11] deposited below the storm wave base[9] (deeper than about 100 m). Neoproterozoic to Early Phanerozoic sedimentary successions suggest that continental deep waters varied from oxic, to ferruginous, to sulfidic, depending on the formation studied[3,5,9–11], and record the geochemistry of spatially and temporally variable local water masses[9,10] instead of the deep, open ocean. Here we present a new proxy for past deep-ocean O_2 concentrations based on $Fe^{3+}/\Sigma Fe$ ratios in ancient hydrothermally altered seafloor basalts.

Today, basalts erupt on the seafloor primarily as pillows and massive flows and are oxidized by oxygenated seawater circulating through oceanic crust. Basalts recovered during deep-sea drilling show that this circulation increases $Fe^{3+}/\Sigma Fe$ ratios from about 0.15 to 0.45 ($\pm$0.15)[17]. Although some of this iron may be oxidized via seawater hydrolysis[17], more than half[17] of all iron oxidation is mediated by O_2. We propose that $Fe^{3+}/\Sigma Fe$ ratios of hydrothermally altered oceanic basalts will be lower in anoxic versus oxygenated deep oceans and could therefore record changes in deep-ocean O_2 concentrations. Although most Mesozoic and older oceanic basalts have been lost via subduction, some are preserved on continents in ophiolite sequences[18]. Following ref. 18, we accept previously identified igneous oceanic crust preserved on continental crust as ophiolites. Finally, ref. 16 defines the modern deep ocean as deeper than 1,200 m—such waters have typical O_2 concentrations[16] of about 180 ± 80 μmol kg^{-1}, are deeper than most oxygen minimum zones[16], and span typical modern oceanic ridge depths (>2,500 m deep[19]).

We compiled $Fe^{3+}/\Sigma Fe$ ratios of ophiolitic basalts from the period 3,503–14 Myr ago (73 ophiolites and 1,085 $Fe^{3+}/\Sigma Fe$ determinations; Supplementary Table 1). Only data from extrusive, subaqueous basalts were compiled because these form the main conduit for seawater circulation through oceanic crust[17]. We additionally compiled $Fe^{3+}/\Sigma Fe$ ratios of Mesozoic–Cenozoic basalts recovered during deep-sea drilling (71 cores and 1,151 $Fe^{3+}/\Sigma Fe$ determinations; Supplementary Table 2).

Figure 1 shows average basalt $Fe^{3+}/\Sigma Fe$ ratios of each ophiolite versus age. We place the following bounds on $Fe^{3+}/\Sigma Fe$ ratios of unaltered oceanic basalts. Oceanic crust in ophiolites derives from a variety of settings including mid-ocean ridges and back-arc basins[18]. We take a lower $Fe^{3+}/\Sigma Fe$ bound for mid-ocean-ridge basalts (whole-rock) of 0.10 (ref. 17) and an upper bound of 0.31 based on Lau Basin back-arc glasses[20]. Mean Archaean (>2,500 Myr ago), Palaeo–Mesoproterozoic (2,500–1,000 Myr ago), and Neoproterozoic (1,000–541 Myr ago) $Fe^{3+}/\Sigma Fe$ ratios are 0.20 ± 0.04 (2 standard errors of the mean; s.e.m.), 0.26 ± 0.02 (2 s.e.m.), and 0.26 ± 0.05 (2 s.e.m.), respectively, and are within the range given for unaltered oceanic crust (0.10–0.31). Statistical pairwise testing (Methods; Extended Data Tables 1 and 2) shows that these means are statistically indistinguishable ($P > 0.05$).

Phanerozoic ophiolite samples (<541 Myr old) have elevated $Fe^{3+}/\Sigma Fe$ ratios relative to Precambrian (>541 Myr old) and unaltered modern basalts and these ratios increase with time. Specifically, $Fe^{3+}/\Sigma Fe$ ratios of Early Palaeozoic ophiolite basalts (541–420 Myr old) average 0.34 ± 0.08 (2 s.e.m.); Late Palaeozoic basalts (420–252 Myr old) average 0.47 ± 0.10 (2 s.e.m.); and Mesozoic–Cenozoic basalts (< 252 Myr old) average 0.58 ± 0.11 (2 s.e.m.). We separate the Early and Late Palaeozoic because some studies argue for a Late Palaeozoic (approximately 420–400 Myr ago[7,10]) oxygenation of the deep ocean. Pairwise testing indicates that Mesozoic–Cenozoic and Late Palaeozoic versus Precambrian means are statistically different ($P < 0.05$). The Early Palaeozoic mean is statistically indistinguishable from the Late Palaeozoic and Neoproterozoic means ($P > 0.05$) but differs from the Mesozoic–Cenozoic, Palaeo–Mesoproterozoic, and Archaean means ($P < 0.05$), supporting the inference of a Phanerozoic increase in ophiolite $Fe^{3+}/\Sigma Fe$ ratios. $Fe^{3+}/\Sigma Fe$ ratios of Mesozoic–Cenozoic drilled oceanic crust average 0.41 ± 0.03 (2 s.e.m.), which we discuss below.

[1]Department of Earth and Planetary Science, University of California, Berkeley, California 94720, USA. [2]Berkeley Geochronology Center, Berkeley, California 94720, USA.

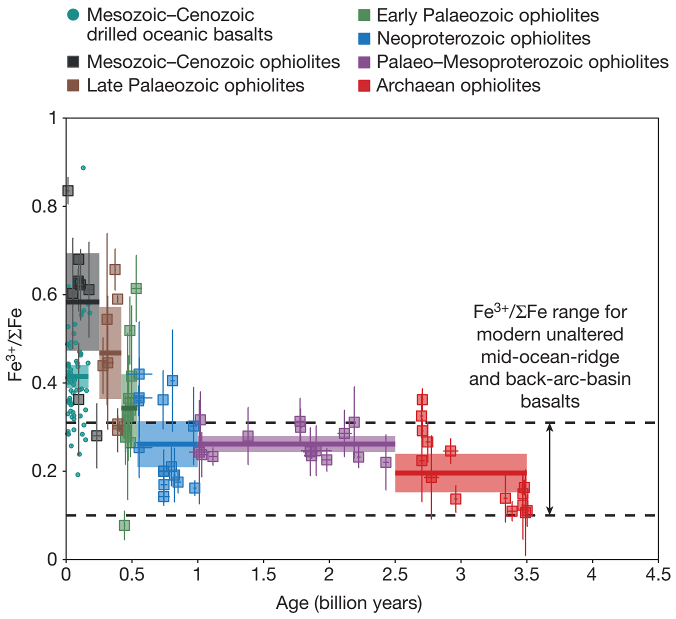


Figure 1 | Average $Fe^{3+}/\Sigma Fe$ ratios of basalts from specific ophiolites and drilled oceanic crust. Solid horizontal lines are sample averages over the given time period. Shading represents 2 s.e.m. uncertainty. $Fe^{3+}/\Sigma Fe$ error bars are 2 s.e.m. Age errors vary depending on the constraints available for each ophiolite (see Supplementary Table 1). Errors for drilled oceanic crust are not given for visual clarity, but are similar to those of the ophiolites. Dotted horizontal lines are the range of $Fe^{3+}/\Sigma Fe$ for modern mid-ocean-ridge and back-arc-basin basalts (see text). Ophiolite basalt $Fe^{3+}/\Sigma Fe$ ratios begin increasing in the Phanerozoic, indicating that the oxygenation of the deep ocean occurred in the Phanerozoic.

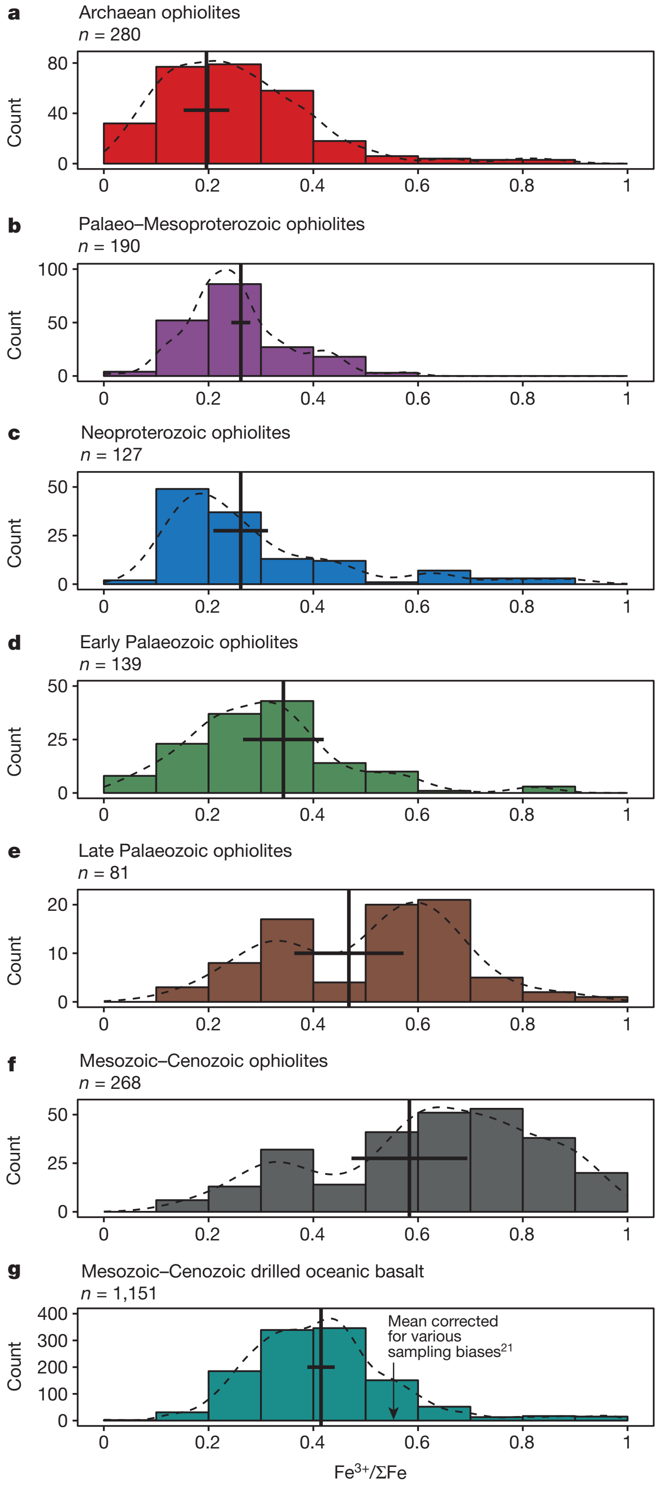


Figure 2 | Histograms of $Fe^{3+}/\Sigma Fe$ ratios from individual samples as a function of time period. a–g, Time periods are the same as those given in Fig. 1. The *y*-axis label 'Count' refers to the number of samples in a given $Fe^{3+}/\Sigma Fe$ bin. Solid black vertical lines are the mean values for a specific time period as given in Fig. 1, with horizontal 2 s.e.m. error bars. *n* is the number of data points. Dotted lines are smoothed distributions (Methods). The difference between Phanerozoic and Precambrian $Fe^{3+}/\Sigma Fe$ ratios is clear both in the average value and in the distributions. **g**, The corrected mean for drilled oceanic crust is taken from ref. 21.

Differences in $Fe^{3+}/\Sigma Fe$ ratios between Precambrian and Phanerozoic ophiolite basalts are apparent in data histograms (Fig. 2). Distributions for Precambrian ophiolites all show peak $Fe^{3+}/\Sigma Fe$ ratios of 0.1–0.3, with a tail to higher values (Fig. 2a–c); we suggest that this tail reflects the recent oxidation of a few samples at Earth's surface. A shift to higher $Fe^{3+}/\Sigma Fe$ ratios emerges in Early Palaeozoic ophiolites (Fig. 2d), as the peak $Fe^{3+}/\Sigma Fe$ ratio rises to 0.3–0.4. This increase continues through the Late Palaeozoic (peak from 0.5–0.7; Fig. 2e) to the Mesozoic–Cenozoic (peak from 0.6–0.8; Fig. 2f).

Late Palaeozoic and Mesozoic–Cenozoic ophiolites have bimodal $Fe^{3+}/\Sigma Fe$ distributions, with a minimum in the range 0.4–0.5. In contrast, the $Fe^{3+}/\Sigma Fe$ distribution of drilled Mesozoic–Cenozoic deep-sea basalts is unimodal, with a maximum (and mean) at about 0.4 (Fig. 2g). We suggest that differences in $Fe^{3+}/\Sigma Fe$ distributions and means between Mesozoic–Cenozoic ophiolites versus drilled oceanic basalts reflect an underestimation of average $Fe^{3+}/\Sigma Fe$ ratios in drilled samples. Specifically, averages of geochemical parameters from drilled oceanic crust can be biased by incomplete core recovery and preferential sampling of pristine or altered samples[17]. Data from Ocean Drilling Program sites 417 and 418 have been used to correct for these biases previously[21]. Mean $Fe^{3+}/\Sigma Fe$ ratios based on averaging all basalt data[17] at these sites range from 0.4–0.45, similar to our drill-core mean of 0.41 (±0.03, 2 s.e.m.). When this mean is calculated using both $Fe^{3+}/\Sigma Fe$ ratios and the relative abundance of various lithologies, this 'weighted' mean increases[21] to 0.56, indistinguishable from the Cenozoic–Mesozoic ophiolite mean (0.58 ± 0.011, 2 s.e.m.). Thus, we argue that mean $Fe^{3+}/\Sigma Fe$ ratios of cored and ophiolite Mesozoic–Cenozoic basalts are in agreement.

The observed Phanerozoic increase in $Fe^{3+}/\Sigma Fe$ ratios could result from several processes including temporal changes in basalt geochemical properties, in metamorphism, or in deep-ocean O_2 concentrations. If basalt iron contents were higher in the past, similar degrees of oxidation could result in smaller changes in $Fe^{3+}/\Sigma Fe$ ratios for older samples. On the basis of $Fe^{3+}/\Sigma Fe$ versus total iron and total iron versus age relationships (Extended Data Fig. 1), we calculate that this process could account for a change in $Fe^{3+}/\Sigma Fe$ ratios of 0.04 ± 0.03 (2σ)

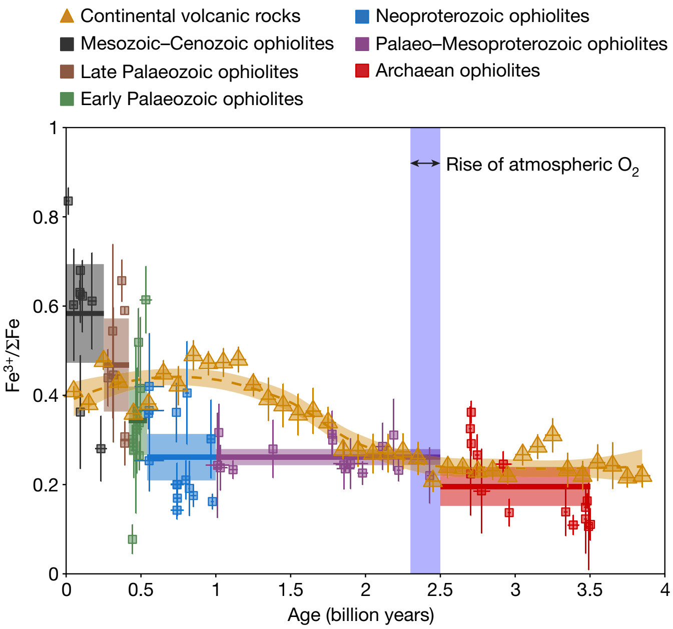


Figure 3 | Comparison of $Fe^{3+}/\Sigma Fe$ ratios from ophiolite basalts versus continental volcanic rocks (including subaerial volcanics). Ophiolite data are the same as in Fig. 1. Continental volcanic $Fe^{3+}/\Sigma Fe$ ratios are means of 100-million-year age bins with 2 s.e.m. error bars[23]. The dotted orange line is a smoothed moving average through the continental volcanic data with a shaded 95% confidence interval (Methods). Continental volcanics increase in $Fe^{3+}/\Sigma Fe$ from 2 to 1.5 billion years ago, more than a billion years before the rise in the ophiolite basalts. This indicates that reductive metamorphism is an unlikely cause of the difference in $Fe^{3+}/\Sigma Fe$ between Phanerozoic and Precambrian ophiolite basalts.

over the past 3.5 billion years in our data. A shift in ophiolite formational environments from mid-ocean ridges to back-arc basins could likewise increase initial basalt $Fe^{3+}/\Sigma Fe$ ratios[22]. However, basalt $Fe^{3+}/\Sigma Fe$ ratios from these environments typically differ[22] by <0.06. Consequently, these processes appear to be unlikely drivers for the >0.3 increase in $Fe^{3+}/\Sigma Fe$ ratios from the Neoproterozoic to Phanerozoic.

Alternatively, lower mean $Fe^{3+}/\Sigma Fe$ ratios of Precambrian versus Phanerozoic samples could reflect preferential metamorphic reduction of Fe^{3+} to Fe^{2+} in Precambrian samples. We consider the late oxidative alteration of samples (for example, occurring on the surface today) to be an unlikely cause of the $Fe^{3+}/\Sigma Fe$ record because this should affect all samples equally and thus does not explain the observed higher $Fe^{3+}/\Sigma Fe$ ratios for Phanerozoic versus Precambrian basalts. For metamorphic reduction to explain the record, it is necessary that $Fe^{3+}/\Sigma Fe$ ratios of Precambrian oceanic basalts first be elevated by oxidation above eruptive values (0.10–0.31) before metamorphic reduction returned them back to their eruptive values, and that this did not occur in Phanerozoic samples. This is possible, because Precambrian ophiolites, owing to their age, are more likely to have experienced metamorphism than Phanerozoic equivalents. We note, however, that samples were selected from systems that retain primary igneous textures (for example, pillows), thus removing highly metamorphosed formations from consideration. Nonetheless, we designed two tests for a metamorphic influence on the $Fe^{3+}/\Sigma Fe$ record:

(1) We examined the relationship between $Fe^{3+}/\Sigma Fe$ and total iron for Precambrian ophiolite basalts (Extended Data Fig. 2); we expect that samples with lower total iron will be more susceptible to metamorphic reduction owing to the lower extent of reaction (that is, the total moles of Fe reduced) needed to reach chemical equilibrium. For these samples, the slope between total iron and $Fe^{3+}/\Sigma Fe$ is -0.003 (± 0.006, 2σ), which is indistinguishable from zero at the 2σ level, and thus is inconsistent with a metamorphic control on the $Fe^{3+}/\Sigma Fe$ record.

(2) We compiled $Fe^{3+}/\Sigma Fe$ ratios from continental volcanic rocks[23] from the past 3,850 Myr (8,335 individual measurements; Fig. 3 and Methods). This compilation includes subaerially erupted rocks, which we expect, following the oxygenation of the atmosphere 2,500–2,300 Myr ago, to be oxidized at the surface by O_2 shortly after eruption. If these rocks also show low $Fe^{3+}/\Sigma Fe$ until the Phanerozoic, it would support the general occurrence of metamorphic iron reduction in Precambrian rocks. This is not the case: from 3,850 Myr ago to 2,000 Myr ago, $Fe^{3+}/\Sigma Fe$ ratios of continental volcanics show mean values similar to those of ophiolite basalts (0.2–0.3), increase from 2,000 Myr ago to 1,500 Myr ago, and then remain approximately constant to the present (Fig. 3). We are unaware of common tectonic processes likely to produce such a specific global and temporal pattern of metamorphism in which continental volcanics preferentially escape reductive metamorphism in the Proterozoic while ophiolites do not. Rather, direct oxidation by atmospheric O_2 following the Proterozoic rise in $p_{O_2,\mathrm{atm}}$ (Fig. 3) provides a simple explanation for the increase in $Fe^{3+}/\Sigma Fe$ of continental volcanics from 2,000 to 1,500 Myr ago.

Consequently, we propose that the Phanerozoic increase in $Fe^{3+}/\Sigma Fe$ ratios of ophiolitic basalts reflects an increase in the hydrothermal flux of O_2 into oceanic crust. One possibility is that a large increase in the flux of already oxygenated seawater through oceanic crust occurred in the Phanerozoic. However, higher fluxes of seawater through oceanic crust have been proposed for the Precambrian relative to the Phanerozoic[24] and hydrothermally altered basalts as old[25] as 3,500 Myr can be found. We therefore consider a large Phanerozoic increase in the flux of seawater through oceanic crust an unlikely driver of the $Fe^{3+}/\Sigma Fe$ record.

Instead, we propose that the Phanerozoic increase in $Fe^{3+}/\Sigma Fe$ ratios of submarine basalts marks the oxygenation of the deep ocean. We now use this record to estimate past deep-ocean O_2 concentrations, employing a model (Methods) that balances the O_2 flux needed to increase basaltic $Fe^{3+}/\Sigma Fe$ ratios from initial (0.10–0.31) to measured values against the minimum seawater O_2 concentration required to supply this flux during hydrothermal alteration of oceanic crust. Modelled O_2 concentrations are minimum estimates because we assume complete reduction of O_2 entering oceanic crust.

Calculated deep-ocean O_2 concentrations are given in Fig. 4a and compared to independent estimates of $p_{O_2,\mathrm{atm}}$ in Fig. 4b. Calculated average Archaean deep-ocean O_2 concentrations are $-3 \pm 18\,\mu\mathrm{mol\,kg^{-1}}$ (2 s.e.m.) and are consistent with the general expectation of an anoxic Archaean deep ocean[2,4,11,12]. Mean calculated Palaeo–Mesoproterozoic and Neoproterozoic O_2 concentrations are $11 \pm 17\,\mu\mathrm{mol\,kg^{-1}}$ (2 s.e.m.) and $11 \pm 20\,\mu\mathrm{mol\,kg^{-1}}$ (2 s.e.m.). Both are within 2 s.e.m. of the anoxic level, broadly compatible with the common view of an anoxic Proterozoic deep ocean[1,2,4–6,11,12], while also allowing for suggestions[8] of low (of the order of micromoles per kilogram of seawater) Proterozoic deep-ocean O_2 concentrations. Calculated average Early to Late Palaeozoic deep-ocean O_2 concentrations are $29 \pm 29\,\mu\mathrm{mol\,kg^{-1}}$ (2 s.e.m.) and $55 \pm 42\,\mu\mathrm{mol\,kg^{-1}}$ (2 s.e.m.). These concentrations compare favourably to estimated minimum O_2 concentrations required for Phanerozoic fauna[7] (more than 15–30 $\mu\mathrm{mol\,kg^{-1}}$). Calculated Mesozoic–Cenozoic deep-ocean O_2 concentrations are $80 \pm 53\,\mu\mathrm{mol\,kg^{-1}}$ (2 s.e.m.), about 45% of modern average deep-ocean O_2 concentrations ($178\,\mu\mathrm{mol\,kg^{-1}}$)[16]. This difference could result if only about 45% of all O_2 that enters oceanic crust is reduced, as opposed to the assumed 100%, which is acceptable given that our calculations are minimum estimates. Alternatively, some estimates of Mesozoic $p_{O_2,\mathrm{atm}}$ are as low as 60%–70% of modern[26], which would support our estimated lower-than-modern Mesozoic deep-ocean O_2 concentrations.

We propose that ophiolite $Fe^{3+}/\Sigma Fe$ ratios provide direct, quantitative constraints on the O_2 content of the deep ocean from the Archaean to the Cenozoic. In particular, they indicate that the deep ocean became oxygenated only in the Phanerozoic and probably not until the late Palaeozoic (<420 Myr; Fig. 4a). This is consistent with some previous studies[7,10,27] (Fig. 4b) but contrasts with proposals for a Precambrian

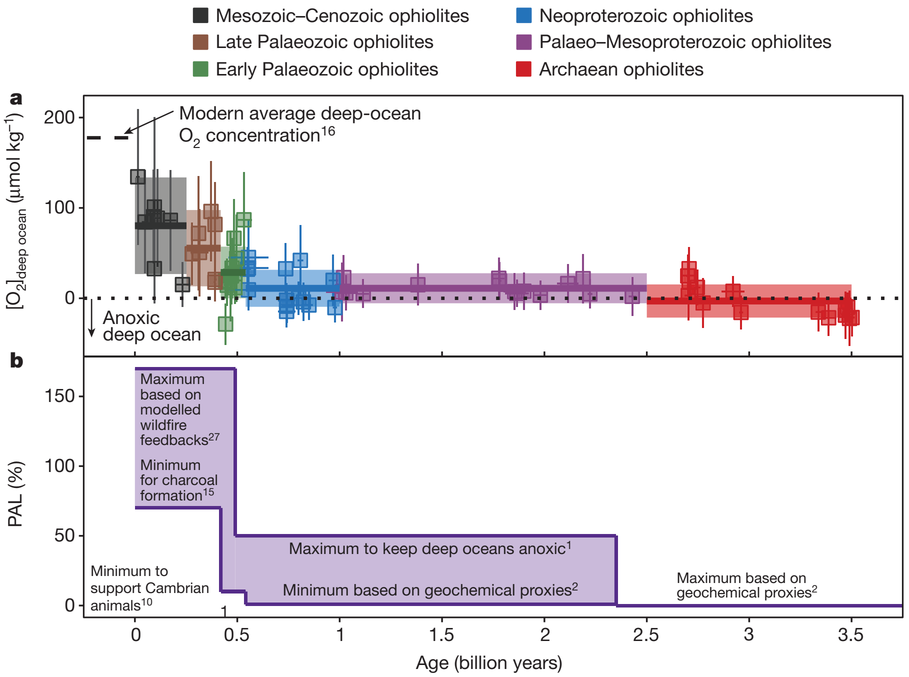


Figure 4 | Calculated deep-ocean O_2 concentrations versus time compared to previous estimates of $p_{O_2,atm}$. **a**, Calculated deep-ocean O_2 concentrations (in micromoles O_2 per kilogram of seawater) based on $Fe^{3+}/\Sigma Fe$ ratios of ophiolite basalts (Methods). Bold horizontal lines are averages for a given time period, with shading representing 2 s.e.m. O_2 concentration error bars are 2 s.e.m. **b**, Ranges of estimated allowable atmospheric O_2 levels are given as present atmospheric levels (PAL), that is, as a percentage of modern levels, and are based on refs 1, 2, 10, 15 and 27. **b** is in part modelled after a figure in ref. 10.

oxygenation of the deep ocean[5,11]. This result is important because it provides direct evidence that the rise in $p_{O_2,atm}$ to levels sufficient to oxygenate the deep ocean (PAL >15%–50%)[1,3] postdates by hundreds of millions of years the Neoproterozoic origination of animals and is thus causally unrelated to this event.

What caused the rise in $p_{O_2,atm}$ that oxygenated the deep ocean and why it postdates the oxygenation of the atmosphere by about 2 billion years is debated, with hypotheses ranging from the radiation of vascular plants[27] to perturbations in global biogeochemical cycles due to Neoproterozoic glaciations[3,5,28]. These hypotheses generally predict rapid (<50 Myr) increases in $p_{O_2,atm}$ to levels sufficient to oxygenate the deep ocean. Additionally, models typically find that Late Palaeozoic $p_{O_2,atm}$ was higher[26] or similar[27] to Mesozoic and Cenozoic $p_{O_2,atm}$. In contrast, the ophiolite $Fe^{3+}/\Sigma Fe$ record suggests that deep-ocean O_2 concentrations and thus $p_{O_2,atm}$ (assuming the two correlate) increased steadily over the entire Phanerozoic. This would indicate that $p_{O_2,atm}$ is regulated by feedbacks that strongly minimize imbalances in O_2 sources and sinks on million-year timescales, preventing large oscillations in $p_{O_2,atm}$. Such a minimization of these imbalances appears to have occurred over the past million years[29].

Finally, the oxygenation of the deep ocean also probably affected the solid Earth. This oxygenation would have led to increased $Fe^{3+}/\Sigma Fe$ ratios in oceanic crust and more oxidized sediments in subducting slabs. This in turn would have increased the subducted flux of oxidized materials to the mantle over the past 540 Myr and caused a progressive oxidation of the mantle over the Phanerozoic. This hypothesis provides a straightforward explanation for invariant estimated upper-mantle oxygen fugacities of modern versus Archaean basalts[30], despite proposals for an increase in mantle oxygen fugacity two to three billion years ago[31] owing to subduction of oxidized Archaean slabs—simply put, such slabs did not become substantially oxidized until the Phanerozoic oxygenation of the deep ocean. This hypothesis is consistent with proposed changes in the uranium isotopic composition of altered oceanic crust and the mantle with time[32] and is testable as follows. If correct, $Fe^{3+}/\Sigma Fe$ ratios should be higher in Phanerozoic than in Precambrian island-arc rocks if arc magmas are oxidized (as some believe[22]) owing to interactions with oxidized fluids derived from oxidized slabs.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 July; accepted 2 November 2017.
Published online 3 January 2018.

1. Canfield, D. E. in *Treatise on Geochemistry* (eds Holland, H. D. & Turekian, K. K.) 197–216 (Elsevier, 2014).
2. Canfield, D. E. The early history of atmospheric oxygen: homage to Robert M. Garrels. *Annu. Rev. Earth Planet. Sci.* **33,** 1–36 (2005).
3. Canfield, D. E., Poulton, S. W. & Narbonne, G. M. Late-Neoproterozoic deep-ocean oxygenation and the rise of animal life. *Science* **315,** 92–95 (2007).
4. Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506,** 307–315 (2014).
5. Sahoo, S. K. *et al.* Ocean oxygenation in the wake of the Marinoan glaciation. *Nature* **489,** 546–549 (2012).
6. Planavsky, N. J. *et al.* Low Mid-Proterozoic atmospheric oxygen levels and the delayed rise of animals. *Science* **346,** 635–638 (2014).
7. Dahl, T. W. *et al.* Devonian rise in atmospheric oxygen correlated to the radiations of terrestrial plants and large predatory fish. *Proc. Natl Acad. Sci. USA* **107,** 17911–17915 (2010).
8. Slack, J., Grenne, T., Bekker, A., Rouxel, O. & Lindberg, P. Suboxic deep seawater in the late Paleoproterozoic: evidence from hematitic chert and iron formation related to seafloor-hydrothermal sulfide deposits, central Arizona, USA. *Earth Planet. Sci. Lett.* **255,** 243–256 (2007).
9. Canfield, D. E. *et al.* Ferruginous conditions dominated later Neoproterozoic deep-water chemistry. *Science* **321,** 949–952 (2008).
10. Sperling, E. A. *et al.* Statistical analysis of iron geochemical data suggests limited late Proterozoic oxygenation. *Nature* **523,** 451–454 (2015).
11. Scott, C. *et al.* Tracing the stepwise oxygenation of the Proterozoic ocean. *Nature* **452,** 456–459 (2008).
12. Canfield, D. E. A new model for Proterozoic ocean chemistry. *Nature* **396,** 450–453 (1998).
13. Blamey, N. J. *et al.* Paradigm shift in determining Neoproterozoic atmospheric oxygen. *Geology* **44,** 651–654 (2016).
14. Scott, A. C. & Glasspool, I. J. The diversification of Paleozoic fire systems and fluctuations in atmospheric oxygen concentration. *Proc. Natl Acad. Sci. USA* **103,** 10861–10865 (2006).

15. Belcher, C. & McElwain, J. Limits for combustion in low O_2 redefine paleoatmospheric predictions for the Mesozoic. *Science* **321,** 1197–1200 (2008).
16. Sarmiento, J. L. & Gruber, N. *Ocean Biogeochemical Dynamics* (Princeton Univ. Press, 2006).
17. Bach, W. & Edwards, K. J. Iron and sulfide oxidation within the basaltic ocean crust: implications for chemolithoautotrophic microbial biomass production. *Geochim. Cosmochim. Acta* **67,** 3871–3887 (2003).
18. Dilek, Y. & Furnes, H. Ophiolite genesis and global tectonics: geochemical and tectonic fingerprinting of ancient oceanic lithosphere. *Geol. Soc. Am. Bull.* **123,** 387–411 (2011).
19. Kasting, J. F. *et al.* Paleoclimates, ocean depth, and the oxygen isotopic composition of seawater. *Earth Planet. Sci. Lett.* **252,** 82–93 (2006).
20. Nilsson, K. & Peach, C. L. Sulfur speciation, oxidation state, and sulfur concentration in backarc magmas. *Geochim. Cosmochim. Acta* **57,** 3807–3813 (1993).
21. Staudigel, H., Plank, T., White, B. & Schmincke, H. U. Geochemical fluxes during seafloor alteration of the basaltic upper oceanic crust: DSDP Sites 417 and 418. *Geophys. Monogr. Ser*. **96,** 19–38 (1996).
22. Kelley, K. A. & Cottrell, E. Water and the oxidation state of subduction zone magmas. *Science* **325,** 605–607 (2009).
23. Keller, C. B. & Schoene, B. Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gyr ago. *Nature* **485,** 490–493 (2012).
24. Halevy, I. & Bachan, A. The geologic history of seawater pH. *Science* **355,** 1069–1071 (2017).
25. Hofmann, A. & Harris, C. Silica alteration zones in the Barberton greenstone belt: a window into subseafloor processes 3.5–3.3 Ga ago. *Chem. Geol.* **257,** 221–239 (2008).
26. Berner, R. A. Phanerozoic atmospheric oxygen: new results using the GEOCARBSULF model. *Am. J. Sci.* **309,** 603–606 (2009).
27. Bergman, N. M., Lenton, T. M. & Watson, A. J. COPSE: a new model of biogeochemical cycling over Phanerozoic time. *Am. J. Sci.* **304,** 397–437 (2004).
28. Laakso, T. A. & Schrag, D. P. Regulation of atmospheric oxygen during the Proterozoic. *Earth Planet. Sci. Lett.* **388,** 81–91 (2014).
29. Stolper, D. A., Bender, M. L., Dreyfus, G. B., Yan, Y. & Higgins, J. A. Pleistocene ice core record of atmospheric O_2 concentrations. *Science* **353,** 1427–1430 (2016).
30. Li, Z.-X. A. & Lee, C.-T. A. The constancy of upper mantle fO_2 through time inferred from V/Sc ratios in basalts. *Earth Planet. Sci. Lett.* **228,** 483–493 (2004).
31. Kump, L. R., Kasting, J. F. & Barley, M. E. Rise of atmospheric oxygen and the "upside-down" Archean mantle. *Geochem. Geophys. Geosyst.* **2,** 1025 (2001).
32. Andersen, M. B. *et al.* The terrestrial uranium isotope cycle. *Nature* **517,** 356–359 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements D.A.S. acknowledges discussions with M. Bender, B. Buffett, D. DePaolo, J. Eiler, S. Finnegan, W. Fischer, J. Higgins, A. Saal and E. Stolper. We thank D. Canfield for advice on a prior draft. C.B.K. acknowledges support from the Ann and Gordon Getty Foundation.

Author Contributions D.A.S. conceived the study and compiled the ophiolite basalt data. C.B.K. provided the continental volcanic data. Both authors contributed to data analysis, modelling and writing.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.A.S. (dstolper@berkeley.edu).

Reviewer Information *Nature* thanks L. Kump and E. Sperling for their contribution to the peer review of this work.

METHODS

Data compilation. Ophiolite basalt $Fe^{3+}/\Sigma Fe$ ratios were compiled from primary studies with reported FeO and Fe_2O_3 values. Data are derived from samples described in previous studies either as subaqueous basalts from ophiolites or preserved oceanic crust. Samples were often identified using previous compilations of Phanerozoic and Precambrian ophiolites[18,33–35]. Studies where Fe_2O_3 values were calculated based on measurements of total FeO contents and assumed FeO/Fe_2O_3 or Fe^{3+}/Fe^{2+} ratios were not used because the Fe_2O_3 values were not independently measured. Additionally, data from Jurassic–Triassic-aged ophiolites in the Brooks Range, Alaska, described in ref. 36, were not used. This study was not used because data were corrected for post-formational oxidative alteration by assuming a relationship between TiO_2 and Fe_2O_3 content. Samples with excess Fe_2O_3 relative to this relationship were assumed to have been altered (though whether this alteration occurred during seafloor alteration or later is not stated) and the excess Fe_2O_3 was converted back to FeO in an attempt to restore the samples to their original igneous compositions. Problematically, which samples were corrected and the originally measured, uncorrected Fe_2O_3 values are not given. As we are interested in a sample's Fe_2O_3 value due to hydrothermal alteration, this study could not be used.

All ophiolite data are given in Supplementary Table 1. We additionally provide in Supplementary Table 1 other major element data, sample descriptions, and age constraints.

For oceanic-basalt data from deep-sea drill cores, $Fe^{3+}/\Sigma Fe$ ratios were compiled from cores taken by the Deep-Sea Drilling Program and its later iterations. Only data for samples older than 10 Myr were used because such samples are thought to have seen sufficient integrated fluid fluxes to reach near-maximum $Fe^{3+}/\Sigma Fe$ ratios[17,37]. Data were derived from original deep-sea drilling reports, as given in Supplementary Table 2. Average values for cores are presented in Fig. 1 and were derived by weighting $Fe^{3+}/\Sigma Fe$ versus depth relative to the depth span defined by the data. This was done to prevent a large number of samples measured from a specific horizon (for example from a vein) from carrying undue weight in the final average. Doing this generally resulted in minimal differences (<0.05) in final $Fe^{3+}/\Sigma Fe$ ratios versus averaging all of the data.

Continental volcanic data were compiled using a database of geochemical measurements[23]. We also included in this database $Fe^{3+}/\Sigma Fe$ ratios from approximately 2-billion-year-old subaerial volcanic rocks recovered from drilling in Fennoscandia[38]. Only igneous rocks with SiO_2 compositions of 40–80 wt% were included in the original[23] compilation (which carries forward to our use of the database). Data were resampled and standard errors of the means for age bins of 100 Myr were calculated following the methodology described in ref. 23. To ensure that the database did not include publications in which $Fe^{3+}/\Sigma Fe$ or FeO/Fe_2O_3 values were assumed or assigned a constant value (in order to calculate an Fe_2O_3 weight per cent value based on an FeO weight per cent measurement for example or vice versa), we removed data from consideration from any publication in which the relative $FeO/(Fe_2O_3+FeO)$ standard deviation (standard deviation/mean) was less than 1%. This removes about 14% of all data, but does not change any trends noticeably.

This continental volcanic data set contains both subaerial and subaqueous volcanic rocks preserved on continental crust. 96% of rocks in the record younger than 100,000 years old are currently found above sea level (and thus probably formed subaerially). How the proportion of subaerial versus subaqueous volcanism changes over time in this database is unknown. However, it has been argued on the basis of abundances of preserved subaqueous versus subaerial large igneous provinces that a large increase in proportion of subaerial versus subaqueous volcanism occurred 2.5 billion years ago[39] (an increase from about 20% to 70%) with relatively stable proportions of subaerial versus subaqueous large igneous province formation from that point onwards. If proportions of subaerial versus subaqueous volcanism for samples from our database mirror the large-igneous-province record, then more than 60% of continental volcanics have formed subaerially since 2.5 billion years ago. If this is correct, then we consider it unlikely that the increase in continental volcanic $Fe^{3+}/\Sigma Fe$ ratio values observed 2 billion years ago (Fig. 3) is linked to changes in proportions of subaerial versus subaqueous volcanism.

Data smoothing. In Fig. 2, a smoothed data distribution (given by the black dotted lines) is presented along with discrete histograms. This smoothed histogram was calculated using the default parameters of the stat_smooth function in the R statistical software package[40]. In Fig. 3, a moving average for the continental volcanic data was given with 95% confidence intervals. It was made using the geom_smooth function with a span of 0.75 in the R statistical software package[40].

Statistical testing. Whether mean $Fe^{3+}/\Sigma Fe$ ratios of various age bins are statistically distinct or not was tested using pairwise testing with a *P*-value cut-off of 0.05, with <0.05 indicating that averages are distinct and >0.05 indicating they are indistinguishable. Two tests were used, the parametric Tukey–HSD test and the non-parametric Wilcoxon test. For the Wilcoxon test, we used the Holm *P*-adjustment method. All tests were performed using functions encoded in the R statistical software package[40]. The *P* values for all pairwise tests are given below in Extended Data Tables 1 and 2.

Modelling dissolved O_2 concentrations. O_2 concentrations were calculated based on measured $Fe^{3+}/\Sigma Fe$ ratios using a mass-balance-based model. This model assumes that a certain volume of crust is produced each year that will have potentially oxidizing seawater flow through it (generally restricted to the top 500 ± 200 m of igneous oceanic crust[17]). It then assumes a certain average weight per cent of iron in the rock produced and uses a specific ophiolite's or average time period's $Fe^{3+}/\Sigma Fe$ ratio value relative to an assumed initial value to calculate the amount of iron oxidized per year in oceanic crust. We assume that the initial $Fe^{3+}/\Sigma Fe$ ratios range from 0.1 to 0.31 (with a uniform distribution); this is the range discussed in the main text for unaltered modern oceanic basalts.

We additionally account for the oxidation of igneous sulphur in oceanic crust. Our compilation lacks measurements of sulphur either in amount or redox state. We assume that all sulphur in the erupted igneous rocks is present as sulphide based on sulphide saturation in mid-ocean-ridge basalts[41]. To calculate the amount of sulphur oxidized in the samples, we assume (for lack of any other constraints) that the ratio of sulphur versus iron oxidation (in moles) observed in modern mid-ocean-ridge basalts has remained constant through time. Thus, from the shift in $Fe^{3+}/\Sigma Fe$ measured in our samples, we can calculate the amount of sulphur that would have been oxidized as well.

We balance the O_2 demand indicated by moles of iron and sulphur oxidized in oceanic crust by an input flux of O_2 into oceanic crust from the deep ocean. This is calculated by multiplying the concentration of O_2 in the deep ocean by the flux of seawater into oceanic crust. As discussed in the main text, in doing this, we assume that all O_2 that is delivered into oceanic crust is consumed via the oxidation of reduced minerals, making our calculations of deep-ocean O_2 concentrations a minimum estimate because some fluids may emerge from oceanic crust with O_2 remaining. Since altered oceanic crust contains secondary minerals that formed in both oxic and anoxic environments[42], it follows that fluids can lose all of their oxygen during circulation. We are unaware of any studies measuring the oxygen concentration of waters flowing out of oceanic crust into the ocean at temperatures <100 °C (hot hydrothermal fluids are anoxic), which would offer a means of testing this assumption.

An additional assumption made is that fluids flowing through oceanic crust today and in the past are sourced from waters that are representative of average deep-ocean water masses. Deep-ocean O_2 concentrations vary as a function of the integrated amount of respiration that has occurred in the deep ocean. Average modern deep-ocean waters ($>1{,}200$ m deep) have O_2 concentrations of about 180 μmol kg^{-1}, and typically are within ±80 μmol kg^{-1} of this number[16]. However, so-called 'oxygen minimum zones' exist in the ocean, where O_2 concentrations can decline[16] to below 10 μmol kg^{-1}. Such water masses are typically restricted to depths[43] shallower than 1,000 m. In contrast, modern spreading centres have typical ocean depths[19] greater than 2,500 m, and thus occur at ocean depths deeper than those at which oxygen minimum zones occur. Instead, fluids flowing through modern oceanic crust are derived from water masses with O_2 concentrations similar to mean deep-ocean ($>1{,}200$ m depth) O_2 concentrations[16] (180 ± 80 μmol kg^{-1}). We note that the depths of oceans may have varied in the past, but the direction of change is unknown and debated[19]. Thus we make what we consider the simplest assumption: that waters flowing through past oceanic basalts track average deep-ocean O_2 concentrations.

The full equation used to balance O_2 consumed versus O_2 supplied to oceanic crust is:

$$\begin{aligned}&\left[\frac{\text{g crust}}{\text{yr}}\right]\times[\text{wt\% Fe}]\times\left[\left(\frac{\text{Fe}^{3+}}{\Sigma\text{Fe}}\right)_{\text{time}}-\left(\frac{\text{Fe}^{3+}}{\Sigma\text{Fe}}\right)_{\text{initial}}\right]\times\left[\frac{1\ \text{mol Fe}}{55.85\ \text{g}}\right]\times\left[\frac{1\,e^{-}}{\text{Fe}^{2+}\ \text{oxidized}}\right]\\&\times\left[\left(\frac{\text{mol S}^{2-}\ \text{oxidized}}{\text{mol Fe}^{2+}\ \text{oxidized}}\right)_{\substack{\text{modern}\\ \text{oceanic crust}}}\times\left(\frac{8\,e^{-}}{\text{S}^{2-}\ \text{oxidized}}\right)+1\right]\times\left[\frac{\text{O}_2\ \text{reduced}}{4\,e^{-}}\right]\\&=\left[\frac{\text{mol O}_2}{\text{kg H}_2\text{O}}\right]_{\substack{\text{deep-ocean}\\ \text{O}_2\ \text{concentration}}}\times\left[\frac{\text{kg H}_2\text{O}}{\text{yr}}\right]_{\substack{\text{flux through}\\ \text{oceanic crust}}}\end{aligned}\tag{1}$$

In equation (1), e^{-} is an electron.

To calculate deep-ocean O_2 concentrations, we solve equation (1) for '$[\text{mol O}_2/\text{kg H}_2\text{O}]_{\text{deep-ocean O}_2\text{ concentration}}$', the first term on the right side of the equation. The term '$(Fe^{3+}/\Sigma Fe)_{time}$' is taken from our data. All other terms are either constants or are taken from the literature.

We take the amount of crust generated per year that sees potentially oxidizing fluids to be 4.0×10^{15} g yr^{-1} with a uniformly distributed error of $\pm 1.8 \times 10^{15}$ g yr^{-1}

as given in ref. 17. In that study[17], the distribution and level of uncertainty reported is not given (for example, whether Gaussian, and, if so, whether the given uncertainty is 1σ or 2σ). Although a stated uncertainty often implies a Gaussian distribution, many quantities that cannot take negative values, such as crustal production rates, are fundamentally incompatible with a Gaussian distribution. To avoid this, we chose a uniform distribution for the error, and assumed that the reported uncertainties above represent the full range of values.

We assume an average weight per cent of Fe of 8% ± 0.7% in oceanic basalts. This is derived from ref. 17, where the error is described as 1σ (and thus has a Gaussian distribution). This value is consistent with other compilations of mid-ocean-ridge-basalt chemical data including that from ref. 44 (7.8% ± 2.9%, 2σ) and ref. 45 (8.1% ± 0.2%, 95% confidence interval).

$Fe^{3+}/\Sigma Fe$ ratios are taken from the ophiolite data. Errors are treated as normally distributed.

The number of moles of S^{2-} oxidized per year in modern oceanic crust is taken as $(1.1 \pm 0.7) \times 10^{11}$ mol yr^{-1}. The number of moles of Fe oxidized per year in modern oceanic crust is taken as $(1.7 \pm 1.2) \times 10^{12}$ mol yr^{-1}. Both values are from ref. 17 and the errors are treated as uniformly distributed for the same reasons as discussed above for crustal production.

We take the values for the flux of seawater to oceanic crust from ref. 46. This study provides modelled estimates of this flux for the past 200 Myr, with ranges from 9×10^{15} to 1.5×10^{16} kilograms of H_2O per year. This overlaps with the range predicted in other calculations[47]. We incorporate the range of estimates for fluid fluxes into our error propagation by randomly sampling a data point from the seawater flux versus time curve (0–200 Myr) in figure 3C of ref. 46. In our error propagation scheme, we take a random number from a uniform, integer distribution from 0 to 200, multiply that number by one million, and then use that number as an age to derive a random seawater flux value from the past 200 Myr.

With all of these various error estimates, we perform a Monte Carlo error propagation scheme to solve equation (1). Specifically, equation (1) is solved for '$[\text{mol } O_2/\text{kg } H_2O]_{\text{deep-ocean } O_2 \text{ concentration}}$' fifty million times for a given input $Fe^{3+}/\Sigma Fe$ ratio (and accompanying error) with numbers for all terms that are not constants drawn from the distributions discussed above.

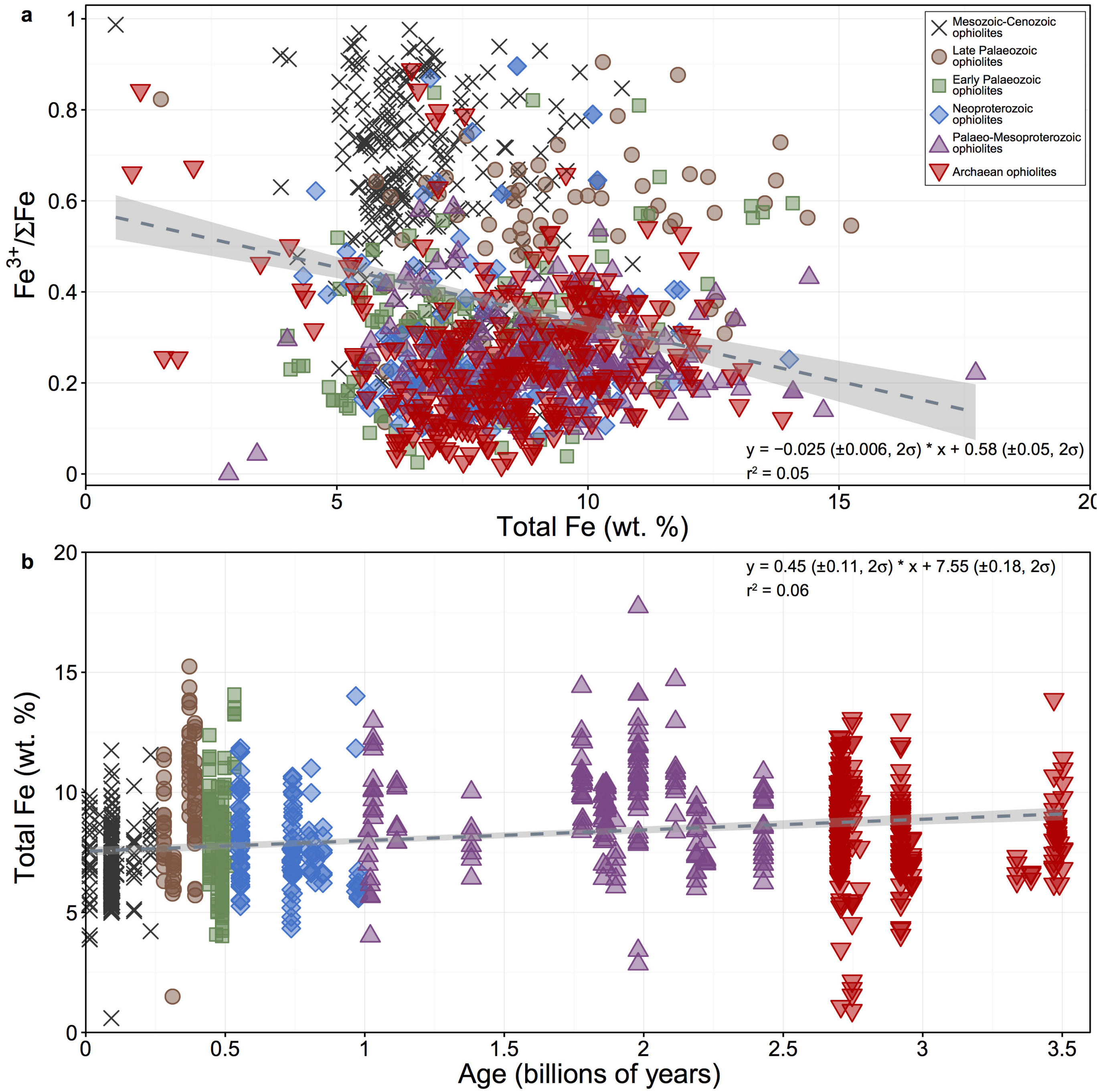
The sensitivity of the model to calculated O_2 concentrations is straightforward because all terms in equation are multiplicative and do not (in our formulation) depend on each other. For example, a twofold increase in Fe weight per cent of basalts (which is on the left side of equation (1)), demands that O_2 concentrations also increase by twofold. Given that the errors on many terms are greater than ±50% relative, we consider our overall error based on the Monte Carlo simulation to be conservative.

We note that a question with this modelling approach is whether ophiolite $Fe^{3+}/\Sigma Fe$ ratios are representative of average oceanic basalt. As discussed in the main text, when mean $Fe^{3+}/\Sigma Fe$ ratios from the Mesozoic–Cenozoic are corrected for sampling biases and poor recovery, they agree with the ophiolite mean for the Mesozoic–Cenozoic (0.56 versus 0.58; see main text). Additionally, ophiolites have long been used to study the hydrothermal alteration of oceanic crust at both low (<100 °C) and high (>100 °C) temperatures in the past[48,49].

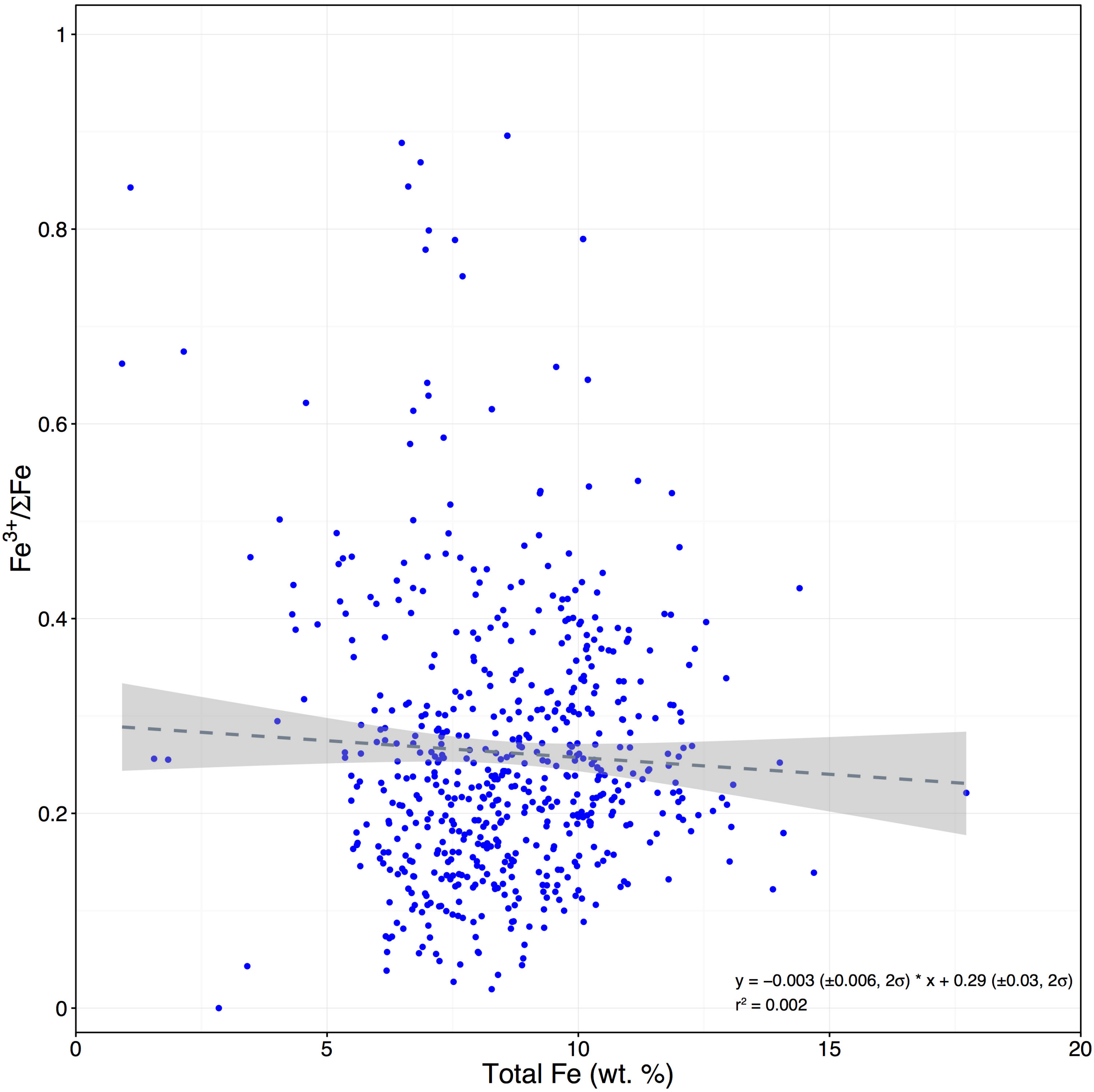
An alternative approach to the model is to normalize all ophiolite data to ophiolites. That is, if we assume, as is typically done, that the Archaean ocean was anoxic[2,4], we can normalize the starting data to the average Archaean ophiolite basalt $Fe^{3+}/\Sigma Fe$ ratio of 0.20 ± 0.04 (2 s.e.m.). For comparison, we have taken the mean $Fe^{3+}/\Sigma Fe$ ratio for unaltered oceanic crust to be 0.205 (the midpoint of 0.10–0.31). As the two are similar (within 0.005 for $Fe^{3+}/\Sigma Fe$ ratios), normalizing to Archaean ophiolites has little effect on the model-derived deep-ocean O_2 concentrations, providing confidence in our approach.

Data availability. All $Fe^{3+}/\Sigma Fe$ ratios compiled for this work are available within the paper and its Supplementary Information.

33. Furnes, H., Dilek, Y. & De Wit, M. Precambrian greenstone sequences represent different ophiolite types. *Gondwana Res.* **27,** 649–685 (2015).
34. Furnes, H., De Wit, M. & Dilek, Y. Four billion years of ophiolites reveal secular trends in oceanic crust formation. *Geosci. Front.* **5,** 571–603 (2014).
35. Lécuyer, C. & Ricard, Y. Long-term fluxes and budget of ferric iron: implication for the redox states of the Earth's mantle and atmosphere. *Earth Planet. Sci. Lett.* **165,** 197–211 (1999).
36. Pallister, J. S., Budahn, J. R. & Murchey, B. L. Pillow basalts of the Angayucham terrane: oceanic plateau and island crust accreted to the Brooks Range. *J. Geophys. Res. Solid Earth* **94,** 15901–15923 (1989).
37. Johnson, H. P. & Semyan, S. W. Age variation in the physical properties of oceanic basalts: implications for crustal formation and evolution. *J. Geophys. Res. Solid Earth* **99,** 3123–3134 (1994).
38. Rybacki, K., Kump, L., Hanski, E. & Melezhik, V. Weathering during the Great Oxidation Event: Fennoscandia, arctic Russia 2.06 Ga ago. *Precambr. Res.* **275,** 513–525 (2016).
39. Kump, L. R. & Barley, M. E. Increased subaerial volcanism and the rise of atmospheric oxygen 2.5 billion years ago. *Nature* **448,** 1033–1036 (2007).
40. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2015); https://www.r-project.org.
41. Wallace, P. & Carmichael, I. S. Sulfur in basaltic magmas. *Geochim. Cosmochim. Acta* **56,** 1863–1874 (1992).
42. Alt, J. C. & Honnorez, J. Alteration of the upper oceanic crust, DSDP site 417: mineralogy and chemistry. *Contrib. Mineral. Petrol.* **87,** 149–169 (1984).
43. Karstensen, J., Stramma, L. & Visbeck, M. Oxygen minimum zones in the eastern tropical Atlantic and Pacific oceans. *Prog. Oceanogr.* **77,** 331–350 (2008).
44. Keller, C. B., Schoene, B., Barboni, M., Samperton, K. M. & Husson, J. M. Volcanic-plutonic parity and the differentiation of the continental crust. *Nature* **523,** 301–307 (2015).
45. Gale, A., Dalton, C. A., Langmuir, C. H., Su, Y. & Schilling, J. G. The mean composition of ocean ridge basalts. *Geochem. Geophys. Geosyst.* **14,** 489–518 (2013).
46. Müller, R., Dutkiewicz, A., Seton, M. & Gaina, C. Seawater chemistry driven by supercontinent assembly, breakup, and dispersal. *Geology* **41,** 907–910 (2013).
47. Elderfield, H. & Schultz, A. Mid-ocean ridge hydrothermal fluxes and the chemical composition of the ocean. *Annu. Rev. Earth Planet. Sci.* **24,** 191–224 (1996).
48. Gregory, R. T. & Taylor, H. P. An oxygen isotope profile in a section of Cretaceous oceanic crust, Samail Ophiolite, Oman: evidence for $\delta^{18}O$ buffering of the oceans by deep ($>$ 5 km) seawater-hydrothermal circulation at mid-ocean ridges. *J. Geophys. Res. Solid Earth* **86,** 2737–2755 (1981).
49. Bickle, M. J. & Teagle, D. A. Strontium alteration in the Troodos ophiolite: implications for fluid fluxes and geochemical transport in mid-ocean ridge hydrothermal systems. *Earth Planet. Sci. Lett.* **113,** 219–237 (1992).



Extended Data Figure 1 | Relationships between Fe^{3+}/ΣFe, total iron, and sample age. a, Fe^{3+}/ΣFe versus total Fe for all ophiolite basalt data. The regression line (dotted grey line) is a linear regression through all data with a 95% confidence interval shaded in grey. **b**, Total Fe versus age for all ophiolite data. The regression line (dotted grey line) is a linear regression through all data with a 95% confidence interval shaded in grey. Combination of the regression lines in **a** and **b** indicates a maximum shift in Fe^{3+}/ΣFe ratios over the past 3.5 billion years of 0.04 ± 0.03 (2σ).



Extended Data Figure 2 | Fe^{3+}/ΣFe versus total Fe for Precambrian ophiolite basalts. The regression line (dotted grey line) is a linear regression through Precambrian ophiolite basalts with a 95% confidence interval shaded in grey. The slope is not distinguishable from 0 at the 2σ level. If secondary metamorphic reduction occurred in these samples, we would expect samples with less Fe to be consistently lower in Fe^{3+}/ΣFe ratios than samples with more Fe. This is not observed, indicating that such metamorphic reduction has not obviously affected the Fe^{3+}/ΣFe ratios of samples.

Extended Data Table 1 | Tukey–HSD pairwise comparison test of average mean ophiolite $Fe^{3+}/\Sigma Fe$ ratios for the given age bins

	Mesozoic-Cenozoic	Late Palaeozoic	Early Palaeozoic	Neo-proterozoic	Palaeo-Meso-proterozoic	Archaean
Mesozoic-Cenozoic	-	0.29	5.1 x 10^{-5}	1.6 x 10^{-8}	1.7 x 10^{-8}	2.2 x 10^{-11}
Late Palaeozoic	0.29	-	0.16	1.2 x 10^{-3}	1.2 x 10^{-3}	9.2 x 10^{-6}
Early Palaeozoic	5.1 x 10^{-5}	0.16	-	0.38	0.39	9.9 x 10^{-3}
Neo-proterozoic	1.6 x 10^{-8}	1.2 x 10^{-3}	0.38	-	1.0	0.56
Palaeo-Meso-proterozoic	1.7 x 10^{-8}	1.2 x 10^{-3}	0.39	1.0	-	0.56
Archaean	2.2 x 10^{-11}	9.2 x 10^{-6}	9.9 x 10^{-3}	0.56	0.56	-

Data used are the mean $Fe^{3+}/\Sigma Fe$ ratios for a given ophiolite. Values are *P* values. Green indicates a *P* value $<$ 0.05 and red a *P* value $>$ 0.05.

Extended Data Table 2 | Wilcoxon pairwise comparison test (with Holm *P*-adjustment correction) of average mean ophiolite $\mathrm{Fe}^{3+}/\Sigma\mathrm{Fe}$ ratios for the given age bins

	Mesozoic-Cenozoic	Late Palaeozoic	Early Palaeozoic	Neo-proterozoic	Palaeo-Meso-proterozoic	Archaean
Mesozoic-Cenozoic	-	0.4010	0.0344	0.0018	0.0004	0.0002
Late Palaeozoic	0.4010	-	0.4010	0.0212	0.0058	0.0018
Early Palaeozoic	0.0344	0.4010	-	0.4010	0.0370	0.0278
Neo-proterozoic	0.0018	0.0212	0.4010	-	0.5125	0.2042
Palaeo-Meso-proterozoic	0.0004	0.0058	0.0370	0.5125	-	0.1581
Archaean	0.0002	0.0018	0.0278	0.2042	0.1581	-

Data used are the mean $\mathrm{Fe}^{3+}/\Sigma\mathrm{Fe}$ ratios for a given ophiolite. Values are *P* values. Green indicates a *P* value < 0.05 and red a *P* value > 0.05.

doi:10.1038/nature25194

Warfare and wildlife declines in Africa's protected areas

Joshua H. Daskin[1]† & Robert M. Pringle[1]

Large-mammal populations are ecological linchpins[1], and their worldwide decline[2] and extinction[3] disrupts many ecosystem functions and services[4]. Reversal of this trend will require an understanding of the determinants of population decline, to enable more accurate predictions of when and where collapses will occur and to guide the development of effective conservation and restoration policies[2,5]. Many correlates of large-mammal declines are known, including low reproductive rates, overhunting, and habitat destruction[2,6,7]. However, persistent uncertainty about the effects of one widespread factor—armed conflict—complicates conservation-planning and priority-setting efforts[5,8]. Case studies have revealed that conflict can have either positive or negative local impacts on wildlife[8–10], but the direction and magnitude of its net effect over large spatiotemporal scales have not previously been quantified[5]. Here we show that conflict frequency predicts the occurrence and severity of population declines among wild large herbivores in African protected areas from 1946 to 2010. Conflict was extensive during this period, occurring in 71% of protected areas, and conflict frequency was the single most important predictor of wildlife population trends among the variables that we analysed. Population trajectories were stable in peacetime, fell significantly below replacement with only slight increases in conflict frequency (one conflict-year per two-to-five decades), and were almost invariably negative in high-conflict sites, both in the full 65-year dataset and in an analysis restricted to recent decades (1989–2010). Yet total population collapse was infrequent, indicating that war-torn faunas can often recover. Human population density was also correlated (positively) with wildlife population trajectories in recent years; however, we found no significant effect, in either timespan, of species body mass, protected-area size, conflict intensity (human fatalities), drought frequency, presence of extractable mineral resources, or various metrics of development and governance. Our results suggest that sustained conservation activity in conflict zones—and rapid interventions following ceasefires—may help to save many at-risk populations and species.

Over the past 70 years, humans have waged war continuously in the world's most biodiverse regions. Between 1950 and 2000, more than 80% of wars overlapped with biodiversity hotspots[9]. In recent decades, the large majority of conflicts have occurred in Africa and Asia—an average of 28 per year since 1989, with no clear indication of slowdown[11]. These continents also support the world's largest numbers of extant large-mammal species (and of those threatened with extinction)[2]. This alignment of warfare and wildlife hotspots might further imperil the world's last remaining assemblages of diverse large-mammal populations, which play important roles in ecosystems and in many local, regional, and national economies[2–4].

Yet there is no consensus as to whether any general, directional relationship exists between armed conflict and biodiversity outcomes[5,8,12,13]. At local scales, both positive and negative effects of war on wildlife have been documented, arising from multiple direct and indirect pathways[8,13]. Negative impacts stem directly from the use of ordnance and chemicals[14], bushmeat hunting by soldiers[15,16], and trade in ivory and other wildlife products to finance military activity[12]; they can also arise indirectly from the weakening of local institutions and the disruption of livelihoods and norms[16]. However, war can also relax pressure on wildlife when people avoid combat zones[10] or are tactically disarmed[8], or when extractive industries decline[8,17]. The net effect of these diverse pathways on wildlife populations has never, to our knowledge, been assessed over continental or multi-decadal scales.

A generalized, large-scale assessment of war's effect on wildlife is needed to help decision-makers to predict the severity and geographical distribution of threats to biodiversity and to develop practical mitigation strategies[13]. This need is perhaps most acute for Africa, where the high frequency, extent, and duration of conflicts[11,18] undermines governance and threatens the livelihoods of rapidly growing human populations, and where large-mammal populations—including many threatened species—have declined sharply[2,19]. A recent modelling study[5] has shown that incorporating conflict risk into protected-area planning improved predicted conservation outcomes throughout Africa; however, the authors noted that the dearth of information about war's ecological impacts remains a major source of uncertainty in such forecasts. Categorical assumptions employed in the absence of pertinent data (for example, that the occurrence of any conflict in a protected area entails total loss of its conservation value[5]) might underestimate returns on conservation investments in volatile regions and cannot account for the likelihood that impacts vary as a function of conflict frequency or intensity[17]. Leaders of conservation organizations have called for increased research on the environmental effects of conflict and the identification of general trends to facilitate mitigation planning in conflict-prone regions[13].

We used spatially and temporally explicit databases[18,20] to quantify the frequency of armed conflict in and around African protected areas since 1946. We first extracted locations for all protected areas covering at least 5 km^2 from the IUCN/UNEP World Database[21], which included 3,585 protected areas from 51 of the 54 countries in Africa. We next mapped conflicts using two datasets that provide the dates and locations of events that caused at least one human fatality and were part of an organized conflict that caused at least 25 fatalities in the year of the event. For 1946–1988, we used PRIO-GRID[18], which delineates conflict zones at annual intervals within 0.5° × 0.5° geographical grid cells. For 1989–2010, we used the Georeferenced Events Database (GED)[20], which provides yearly minimum convex polygons encompassing conflict locations along with associated estimates of human fatalities. We translated GED polygons into PRIO-GRID structure using a spatial join in ArcMap 10.0, enabling us to quantify the number of years of conflict in each grid cell from 1946 to 2010; we then calculated the mean number of conflict-years for each protected area by averaging across all grid cells that wholly or partially overlapped that protected area.

[1] Department of Ecology & Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA. †Present Address: Department of Ecology & Evolutionary Biology and Yale Institute for Biospheric Studies, Yale University, New Haven, Connecticut 06520, USA.

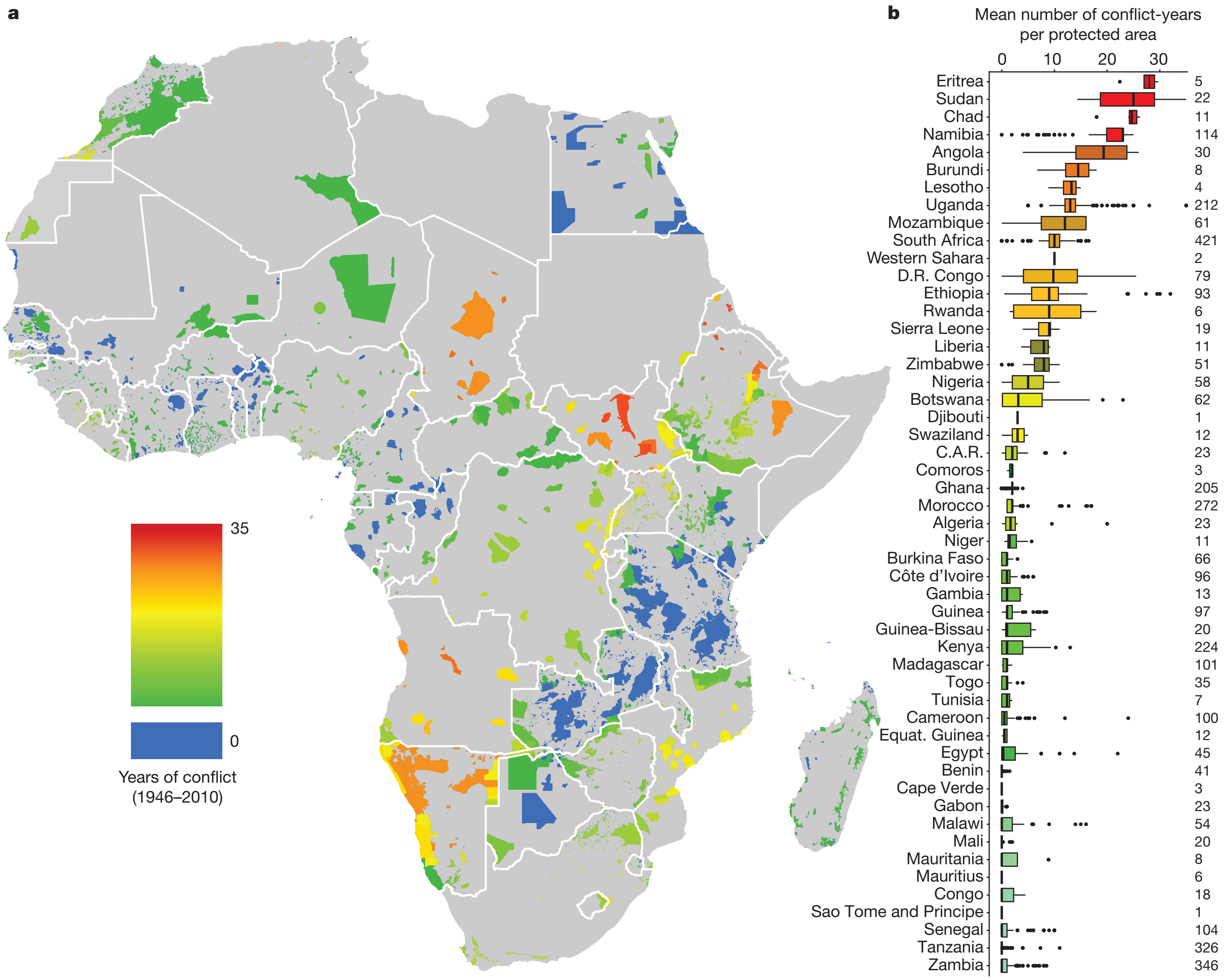


Figure 1 | Geographical distribution and frequency of armed conflict in African protected areas, 1946–2010. a, Number of conflict-years in each protected area; colours indicate average value across all grid-cells overlapping the protected area. **b**, Mean conflict-years per protected area in each country. Boxes, inter-quartile ranges; vertical lines, medians; whiskers, 1.5× inter-quartile range from the median; dots, outliers. Total number of protected areas per country, from the World Database of Protected Areas[21], is shown on the right; statistical analyses of correlations between conflict and wildlife population trajectories were based on the subset of these protected areas for which adequate wildlife data were available. Sudan and South Sudan are distinguished in **a** but combined in **b**; two outlying island nations, Cape Verde and Mauritius, are omitted from **a** but included in **b**. Map created in ArcGIS and R using open-access country-border data from the Global Administrative Areas database (http://gadm.org). C.A.R., Central African Republic; D.R. Congo, Democratic Republic of Congo; Equat. Guinea, Equatorial Guinea.

The mean number of conflict-years in African protected areas during this interval ranged from 0 to 35 (Fig. 1). Continent-wide, 71% of protected areas overlapped at least partially with one or more conflict, and 25% had at least 9 mean conflict-years. For the period 1989–2010 (covered exclusively by the GED dataset), the number of conflict-years ranged from 0 to 19, and 42% of protected areas overlapped with at least one conflict (Extended Data Fig. 1).

We compiled fine-scale data on wildlife population densities by systematically reviewing academic and grey literature (Extended Data Table 1) and extracting data on the abundance of large ($\geq$5 kg) herbivorous mammal populations that had been quantified at multiple times within the same protected area between 1946 and 2010. We filtered these records using stringent quality-control criteria to ensure that the original abundance estimates were made using repeatable methods and that paired estimates for each population were comparable. The final dataset included contrasts for 253 populations, representing 36 species (body-size range: 5–3,825 kg) from 126 protected areas in 19 countries (Supplementary Tables 1, 2). We calculated population trajectories as the annualized finite rate of population change, $\lambda = (D_2/D_1)^{1/(Y_2-Y_1)}$, where Y_1 and Y_2 are the years in which densities D_1 and D_2 were estimated, respectively. Hence, $\lambda = 1$ indicates a stable population, $\lambda > 1$ a growing population, and $\lambda < 1$ a declining population. When multiple λ estimates were available for a given population, we selected just one of them for analysis, using standardized criteria designed to maximize data quality and minimize pseudoreplication; a bootstrapped sensitivity analysis showed that our results were robust to this filtering procedure (Extended Data Table 2).

We used linear regression with model-selection and model-averaging[22] to identify correlates of λ and to calculate parameter estimates for those predictors. For this analysis, we defined the conflict frequency associated with each λ by calculating the proportion of conflict-years between Y_1 and Y_2 in each grid cell and then averaging these proportions across all grid cells overlapping the protected area. Conflict frequency ranged from 0 to 1, and 37% of λ estimates had conflict frequencies greater than 0. In addition to conflict frequency, we tested nine other biological, environmental, geographical, and

Table 1 | Model-averaged parameter estimates for predictors of λ

1946–2010	β	s.e.	β_{LCL}	β_{UCL}	RVI
Intercept	0.00	0.00	N/A	N/A	N/A
Conflict frequency*	−0.06	0.02	−0.09	−0.02	0.98
Protected-area size	0.02	0.02	−0.004	0.07	0.62
Body mass	0.02	0.02	−0.01	0.06	0.51
1989–2010	β	s.e.	β_{LCL}	β_{UCL}	RVI
Intercept	0.00	0.00	N/A	N/A	N/A
Conflict frequency*	−0.08	0.02	−0.12	−0.03	0.98
Human population density*	0.07	0.02	0.03	0.12	0.97
Percentage of urban area	0.02	0.02	−0.01	0.08	0.53
Drought frequency	0.02	0.02	−0.01	0.08	0.48
Resource presence	−0.01	0.02	−0.07	0.02	0.45
Body mass	0.01	0.02	−0.02	0.07	0.44
Protected-area size	0.01	0.02	−0.03	0.07	0.34
Corruption Perceptions Index	0.004	0.01	−0.03	0.06	0.30
Travel time to nearest urban centre	−0.003	0.01	−0.06	0.03	0.28
Conflict intensity	−0.0007	0.01	−0.05	0.04	0.26

Standardized and centred model-averaged parameter estimates (β), standard errors (s.e.), upper (β_{LCL}) and lower (β_{UCL}) 95% confidence limits, and RVI for each predictor for 1946–2010 ($n = 253$ λ estimates) and 1989–2010 ($n = 172$ λ estimates). Statistically significant predictors (those with confidence limits not overlapping 0) are indicated by asterisks.

governance metrics that prior research has suggested might influence wildlife populations: conflict intensity (number of conflict-related human fatalities per km^2 per year, from the GED[23]); body mass (kg), a correlate of life-history traits associated with extinction proneness[6]; protected-area size (km^2), expected to be positively correlated with population persistence[24]; human population density (HPD, individuals per km^2), a proxy for anthropogenic pressure[24]; the Corruption Perceptions Index, a national governance metric that has been linked with wildlife declines[25]; percentage of urban area within grid cells, which might influence λ if protected areas near cities are better-funded and better-managed (or, alternatively, more degraded[23]); travel time to nearest urban centre (minutes), which differs from percentage of urban area because it includes distances to cities that lie outside focal grid cells[23]; drought frequency (proportion of months per year in which the SPEI drought index[26] was $\geq$1.5 s.d. below the long-term average), expected to negatively affect λ; and presence of extractable mineral resources, which can exacerbate conflict and reduce biodiversity[13,17].

Only three of these variables (conflict frequency, body mass, and protected-area size) were available for the entire period 1946–2010. We therefore conducted one analysis for this full dataset of 253 λ estimates using these three predictors, and another for a restricted dataset of 172 estimates from 1989 to 2010 (Extended Data Tables 3, 4); the latter used conflict data from a single source (GED) along with all ten predictors. For each interval, we specified a priori a candidate set of models that included all possible additive combinations of available variables along with a null intercept-only model[22]. We ranked models for each interval using Akaike's information criterion (AIC$_c$) and used Akaike weights (w_i, the likelihood of a model's being the best in the candidate set) to identify the 95% confidence set (that is, all highest-ranked models with summed $w_i \leq 0.95$)[22]. For each predictor in the 95% set, we averaged the standardized parameter estimates across models; predictors were deemed statistically significant when the 95% confidence intervals of their averaged coefficients did not overlap zero[22]. We also calculated the overall relative variable importance (RVI)[22] for each predictor in each interval by summing w_i for all models in which that predictor occurred.

We examined the residuals of the best-fitting models, verifying that the assumptions of regression were not violated (Extended Data Figs 2, 3). We also found no marked spatial autocorrelation in the residuals of these models and ascertained that our main results were robust to potential pseudoreplication arising from inclusion of λ estimates for more than one species within the same protected area (Extended Data Fig. 4 and Extended Data Table 5). For full details of our data, statistical approach, model validation, and sensitivity analyses, see Methods, Extended Data Figs 2–4 and Extended Data Tables 1–6, Supplementary Tables 1–5, and Supplementary Data 1, 2.

Wildlife population trajectories declined as conflict frequency increased (Fig. 2). In both intervals, conflict frequency was included in the single best-fitting model (Extended Data Tables 3, 4), had significantly negative model-averaged coefficients, and had the highest RVI (0.98, indicating that all models with substantial empirical support included conflict frequency; Table 1). In peaceful protected areas (zero conflict frequency), individual population trajectories were variable but collectively averaged very close to the replacement value of 1.0 ($\lambda = 0.99$ for both intervals), indicating that populations were generally stable in the absence of conflict. However, even low levels of conflict reduced λ to significantly below replacement: upper 95% confidence limits for fitted values of λ in the best-fitting models intersected $\lambda = 1.0$ when conflict frequency exceeded 0.02 and 0.05 (equivalent to 1 conflict-year per 2 and 5 decades) for the full and restricted intervals, respectively (Fig. 2).

For 1946–2010, no other variable had a statistically significant model-averaged effect on λ (Table 1). For 1989–2010, HPD had a significantly positive model-averaged coefficient (and RVI = 0.97), suggesting that wildlife fared better in more densely populated areas. All ten predictors were retained in the 95% confidence set of models for this interval, but none had a significant model-averaged coefficient (Table 1). Thus, although the best-fitting models explained a limited amount of the overall variance in λ ($R^2 = 0.05$–0.10)—probably reflecting the many other factors that cause variation in wildlife population dynamics, including species interactions, disease, environmental perturbations, and interspecific differences in sensitivity to such factors—conflict frequency consistently predicted wildlife declines in both intervals.

The observed association between HPD and λ from 1989 to 2010, although counterintuitive and contrary to some prior findings[7,24], is consistent with other large-scale studies that have found positive correlations between HPD and large-mammal population metrics, at least over a subset of body sizes or HPD values[6,27]. We were unable to evaluate the mechanism underlying this association (which does not necessarily imply a direct causal link), but it cannot be explained by collinearity between human density and conflict, because HPD was positively correlated with both conflict frequency (linear regression, $R^2 = 0.12$, $F_{1,170} = 22.7$, $P < 0.0001$) and conflict intensity (linear regression, $R^2 = 0.14$, $F_{1,170} = 28.5$, $P < 0.0001$) in our dataset. Similarly, we found no evidence for one plausible hypothesis for the correlation between HPD and λ—that wildlife in protected areas closer to cities benefit from more attentive management or less illegal hunting—in that neither of our 'urbanization' metrics significantly predicted λ (Table 1). A more direct test of this hypothesis, using data on funding, visitation, poaching, or management regime, would help to clarify what is probably a nuanced relationship between human settlement patterns and wildlife population trends in protected areas.

Extinction risk increases with body size across all mammals[4,7,28], yet we detected no significant effect of body mass on λ in either interval (Table 1). This is probably in part because our data included only herbivores with a body mass of at least 5 kg, with just 12 records for species weighing less than 40 kg (reflecting the general paucity of data for small species, which are harder to count reliably); most species in this medium-to-large size range are targeted by hunters[6] and are similarly likely to be threatened[28]. However, macro-ecological studies caution that simple negative relationships between body size and population persistence cannot be assumed: population trends for mammals in protected areas worldwide tend to increase with body mass[29], and the relationship between size and extinction risk varies both geographically[7] and across the body-size range included in our study[6]. Likewise, although larger protected areas are typically expected to mitigate extinction risk[24], we found no effect of protected-area size on λ, and again our result is not unique: a recent study of population trends in mammals and birds worldwide also found no marked effect

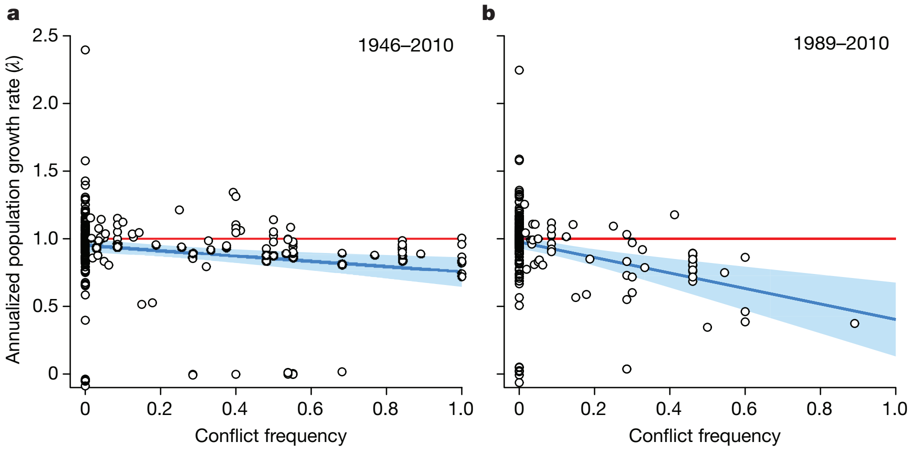


Figure 2 | Conditional regression plots of annualized wildlife population growth rate (λ) as a function of conflict frequency. **a**, **b**, Blue lines represent the effect of conflict frequency on λ in the best-fitting multiple-regression models for 1946–2010 (**a**, $n = 253$) and 1989–2010 (**b**, $n = 172$). Points are partial residuals, conditional on the median values of the other variables in each model (body mass, 592.7 kg; protected-area size, 3,245 km^2 for 1946–2010; HPD, 19.65 per km^2; drought frequency, 0.032; percentage of urban area, 0 for 1989–2010). Shading represents 95% confidence bands for fitted values of λ; red lines at $\lambda = 1$ indicate population-replacement level.

of protected-area size or shape, potentially reflecting context-specific variability in management or socioeconomic factors[29].

Conflict intensity was the least important predictor of λ in the 1989–2010 analysis (Table 1), and a pairwise scatterplot revealed no relationship between these two variables. The greater explanatory power of conflict frequency relative to conflict intensity suggests that the mere occurrence of conflict, irrespective of its human death toll, was sufficient to diminish wildlife populations. Accordingly, we hypothesize that the effects of socioeconomic upheaval and livelihood disruption associated with conflict outweigh the direct effects of military activity. Testing this conjecture will require further research and the cultivation of new data sources; however, it is at least superficially consistent with previous work showing that of 24 mechanisms by which armed conflict is known to affect wildlife, the eight most frequently reported (and 86% of all those cited in case studies) were ‘non-tactical’ pathways involving institutional decay, displacement of people, and altered economies[8].

Although individual conflicts can have either positive or negative impacts on wildlife populations[8,13], we show that the overarching trend is negative, and that even low-grade, infrequent conflict is sufficient to drop population trajectories below replacement. However, outcomes were less severe than those assumed in a pioneering 2016 effort to incorporate conflict risk into systematic conservation planning[5]: population-extinction events were infrequent, even at sites with high conflict frequencies (Fig. 2), suggesting that post-conflict recovery will typically be possible. Thus, whereas we agree that conflict risk should be used to inform conservation investments[5], our results suggest that a risk-tolerant approach may often be warranted. It has been recommended[13] that conservation organizations should conduct pre-conflict contingency planning, maintain a presence during conflict, and use post-conflict windows of opportunity to enact biodiversity-friendly policy change in collaboration with humanitarian relief organizations. Although we are unaware of any published assessment of the efficacy of these approaches (which represents a promising avenue for future inquiry), anecdotal evidence provides grounds for optimism: post-conflict rehabilitation initiatives in Mozambique’s Gorongosa National Park and Rwanda’s Akagera National Park have successfully linked poverty alleviation and human development with improved protected-area administration, ranger training, and wildlife monitoring to enhance conservation outcomes[30].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 November 2016; accepted 5 December 2017.
Published online 10 January 2018.

1. Charles-Dominique, T. *et al.* Spiny plants, mammal browsers, and the origin of African savannas. *Proc. Natl Acad. Sci. USA* **113,** E5572–E5579 (2016).
2. Ripple, W. J. *et al.* Collapse of the world’s largest herbivores. *Sci. Adv.* **1,** e1400103 (2015).
3. Ceballos, G. & Ehrlich, P. R. Mammal population losses and the extinction crisis. *Science* **296,** 904–907 (2002).
4. Dirzo, R. *et al.* Defaunation in the Anthropocene. *Science* **345,** 401–406 (2014).
5. Hammill, E., Tulloch, A. I. T., Possingham, H. P., Strange, N. & Wilson, K. A. Factoring attitudes towards armed conflict risk into selection of protected areas for conservation. *Nat. Commun.* **7,** 11042 (2016).
6. Cardillo, M. *et al.* Multiple causes of high extinction risk in large mammal species. *Science* **309,** 1239–1241 (2005).
7. Fritz, S. A., Bininda-Emonds, O. R. & Purvis, A. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecol. Lett.* **12,** 538–549 (2009).
8. Gaynor, K. M. *et al.* War and wildlife: linking armed conflict to conservation. *Front. Ecol. Environ.* **14,** 533–542 (2016).
9. Hanson, T. *et al.* Warfare in biodiversity hotspots. *Conserv. Biol.* **23,** 578–587 (2009).
10. Hallagan, J. B. Elephants and the war in Zimbabwe. *Oryx* **16,** 161–164 (1981).
11. Pettersson, T. & Wallensteen, P. Armed conflicts, 1946–2014. *J. Peace Res.* **52,** 536–550 (2015).
12. Dudley, J. P., Ginsberg, J. R., Plumptre, A. J., Hart, J. A. & Campos, L. C. Effects of war and civil strife on wildlife and wildlife habitats. *Conserv. Biol.* **16,** 319–329 (2002).
13. Oglethorpe, J., Ham, R., Shambaugh, J. & van der Linde, H. in *Conserving the Peace: Resources, Livelihoods, and Security* (eds Matthew, R. *et al.*) 361–383 (IISD and IUCN, 2002).
14. Orians, G. H. & Pfeiffer, E. W. Ecological effects of the war in Vietnam. *Science* **168,** 544–554 (1970).
15. Beyers, R. L. *et al.* Resource wars and conflict ivory: the impact of civil conflict on elephants in the Democratic Republic of Congo - the case of the Okapi Reserve. *PLoS One* **6,** e27129 (2011).
16. de Merode, E. *et al.* The impact of armed conflict on protected-area efficacy in Central Africa. *Biol. Lett.* **3,** 299–301 (2007).
17. Butsic, V., Baumann, M., Shortland, A., Walker, S. & Kuemmerle, T. Conservation and conflict in the Democratic Republic of Congo: The impacts of warfare, mining, and protected areas on deforestation. *Biol. Conserv.* **191,** 266–273 (2015).
18. Tollefsen, A. F., Strand, H. & Buhaug, H. PRIO-GRID: A unified spatial data structure. *J. Peace Res.* **49,** 363–374 (2012).
19. Craigie, I. D. *et al.* Large mammal population declines in Africa’s protected areas. *Biol. Conserv.* **143,** 2221–2228 (2010).
20. Sundberg, R. & Melander, E. Introducing the UCDP Georeferenced Event Dataset. *J. Peace Res.* **50,** 523–532 (2013).
21. IUCN & UNEP-WCMC. The World Database on Protected Areas (WDPA) https://www.protectedplanet.net/ (2014).
22. Anderson, D. R. *Model Based Inference in the Life Sciences: A Primer on Evidence* (Springer, 2008).
23. McDonald, R. I., Kareiva, P. & Forman, R. T. The implications of current and future urbanization for global protected areas and biodiversity conservation. *Biol. Conserv.* **141,** 1695–1703 (2008).
24. Brashares, J. S., Arcese, P. & Sam, M. K. Human demography and reserve size predict wildlife extinction in West Africa. *Proc. Biol. Sci.* **268,** 2473–2478 (2001).
25. Smith, R. J., Muir, R. D. J., Walpole, M. J., Balmford, A. & Leader-Williams, N. Governance and the loss of biodiversity. *Nature* **426,** 67–70 (2003).

26. Beguería, S., Vicente-Serrano, S. M. & Angulo-Martínez, M. A multiscalar global drought dataset: the SPEIbase: A new gridded product for the analysis of drought variability and impacts. *Bull. Am. Meteorol. Soc.* **91,** 1351–1356 (2010).
27. Ogutu, J. O. *et al.* Extreme wildlife declines and concurrent increase in livestock numbers in Kenya: What are the causes? *PLoS One* **11,** e0163249 (2016).
28. Davidson, A. D., Hamilton, M. J., Boyer, A. G., Brown, J. H. & Ceballos, G. Multiple ecological pathways to extinction in mammals. *Proc. Natl Acad. Sci. USA* **106,** 10702–10705 (2009).
29. Barnes, M. D. *et al.* Wildlife population trends in protected areas predicted by national socio-economic metrics and body size. *Nat. Commun.* **7,** 12747 (2016).
30. Pringle, R. M. Upgrading protected areas to conserve wild biodiversity. *Nature* **546,** 91–99 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank E. Angus, C. Baker, C. Buoncore, J. Castillo Vardaro, B. Lin, A. Tilman, and the Princeton University libraries staff for assistance with data collection and analysis; D. Wilcove, S. Pacala, S. Morris, S. Budischak, J. Socolar, K. Gaynor, J. Edmond, T. Coverdale, R. Long, and U. Srinivasan provided comments. I. Craigie provided access to raw data from ref. 19, and we acknowledge the use of publically available data sources, especially from R. East, the IUCN African Elephant Specialist Group, the World Database of Protected Areas, PRIO-GRID, and GED. This work is a product of US NSF DDIG grant DEB-1501306 to J.H.D. and R.M.P. Additional support was provided by NSF DEB-1355122, DEB-1457697, and the Princeton Environmental Institute.

Author Contributions J.H.D. and R.M.P. conceived the research. J.H.D. designed the study, collected and curated the data, and performed all statistical analyses. R.M.P. provided input on study design, data collection, analysis, and interpretation. Both authors wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.H.D. (joshua.daskin@yale.edu) or R.M.P. (rpringle@princeton.edu).

Reviewer Information *Nature* thanks I. Craigie, O. Schmitz, A. Shortland and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Obtaining population-density estimates from the literature. We systematically compiled data on large-mammal population densities by searching peer-reviewed publications, databases, reports, and other literature. We began by searching ISI Web of Science using the terms in Extended Data Table 1. This yielded 3,113 references, of which we deemed 213 suitable for in-depth review. We extracted data directly from these studies and followed citation trails to additional relevant sources and data. Obscure and out-of-print sources were obtained with help from Princeton University's interlibrary-loan service or by contacting the authors directly. In total, we reviewed 479 sources.

We focused exclusively on wild large herbivores (>5 kg). We included only estimates from contiguous populations of single species, derived over ≤ 2 years between 1946 and 2010, with geographical extent overlapping at least one protected area registered in the IUCN World Database of Protected Areas[21]. We used only estimates from aerial- or ground-based sample or total counts, or from counts of uniquely identifiable individuals. We excluded estimates based on expert knowledge alone and those for which we were unable to verify that count methodology satisfied our criteria. Finally, we included only estimates that specified the area over which the estimate applied; we used this area to calculate density when only raw abundance was reported.

When more than one source reported estimates for a given population in the same year—or when sources included population estimates specific to particular seasons or designated locations within a protected area—we retained each estimate, pending additional filtering. When one source provided more than one estimate for the same population in a given year using identical methods without specifying differences in season or habitat (for example, when repeated counts were conducted to enhance replication and accuracy), we averaged the estimates. When one source reported more than one estimate for the same population in a given year using different methods, we preferentially retained total counts over sample counts. This process netted 3,834 density estimates from 114 of the 479 sources that we reviewed. Summary data for all protected areas, populations, and species are provided in Supplementary Tables 1 and 2.

Calculation of wildlife population trajectories and quality-control criteria. We calculated annualized finite rates of population change, λ (refs 31, 32), for populations with estimates from two time points. To help ensure comparability of estimates from different time points, we used only pairs made using the same field method. To avoid bias stemming from differential sampling coverage, we excluded pairs for which the smaller areal extent was $<25\%$ of the larger. We further restricted analyses to pairs with ≥ 3 years between estimates and a starting population size of more than 50 individuals, to limit the influence of demographic and environmental stochasticity in small populations over short durations.

For the subset of populations with estimates from more than two time points, we selected just one pair of estimates by applying a sequential series of filters (designed to both maximize data quality and minimize statistical non-independence and pseudoreplication). These filters were applied separately in each of the two time periods, 1946–2010 and 1989–2010. First, because ecological data from conflict regions are scarce, we preferentially selected pairs associated with the greatest conflict frequency. If more than two estimates remained after this step, we chose the pair that spanned the longest temporal interval. These two steps eliminated 98% of redundancies. We next selected pairs that were most similar in the seasonality and areal extent (in that order) of the original surveys. Finally, we conservatively selected the pair with a population trajectory closest to replacement. Although this duplicate-removal procedure was guided by data-quality considerations and did not use any arbitrary rules, we nonetheless wanted to test the sensitivity of our results to this process. We therefore used a bootstrap procedure to randomly draw single λ values from those available, which yielded comparable parameter estimates (see 'Sensitivity analyses').

Predictor variables. The predictor variables in our analyses were selected (a) on the basis of a priori hypotheses[22] about factors that we considered (or that prior work has shown to be) likely to influence large-herbivore trajectories and (b) to be minimally redundant with other predictors. We describe each of these predictors, and the rationale for their inclusion, below.

Conflict data were obtained from the Peace Research Institute Oslo's PRIO-GRID v.1.01[18] (for 1946–1988) and the Uppsala Conflict Data Program's Georeferenced Events Database (GED) polygon layer v.1.1[20,33] (for 1989–2010). Both datasets were assembled by teams of geocoders based on organized violence reported in news media, and were developed to enable spatially explicit, fine-scale quantitative study of armed conflict[34–37]. The spatial resolution of the PRIO-GRID and GED datasets (0.5° × 0.5°, equivalent to 55 km × 55 km at the equator, smaller at higher latitudes) is appropriate for analysing the effects of conflict on individual wildlife populations in protected areas that wholly or partially overlap with affected grid cells. Such gridded data unavoidably include spillover across national and other political boundaries. However, transboundary impacts of conflict on biodiversity occur frequently[38–40] and it is therefore appropriate to use conflict metrics that incorporate them.

We chose PRIO-GRID and the GED over another commonly used spatially explicit conflict dataset, the Armed Conflict Location & Event Data Project (ACLED)[41], for two reasons. First, PRIO-GRID and the GED together cover a much longer temporal record (1946–2010)[20] than does ACLED (1997 onwards). This allowed us to include data from more sites and countries, and hence a larger sample size, which was necessary owing to both the general scarcity of wildlife data in conflict areas and the variability inherent in such data. Second, previous work has found that the GED conflict data contain fewer errors than those in ACLED[42].

Our measure of conflict frequency—the mean proportion of years in which conflict occurred during the interval over which λ was calculated—ranged from 0 (conflict-free) to 1 (conflict in all years and grid cells); thus, for example, a conflict frequency of 0.50 could indicate either that half of the grid cells in a protected area were in conflict each year, or that all cells were in conflict for half the years. Because conflict frequency incorporates even relatively minor skirmishes (a minimum of one battle death in the context of a larger conflict), we also devised an index of conflict intensity by extracting the annual number of fatalities per km^2 in each polygon from the GED dataset[20], assigning this value to all grid cells overlapping the polygon, and averaging across grid cells for the protected areas and years associated with each λ estimate.

We included protected-area size because previous work has shown that extinction risk is greater in smaller reserves[24,43]. We obtained protected-area sizes from the World Database of Protected Areas website (https://protectedplanet.net)[21]. When more than one protected area occurred within the area encompassed by a given density estimate, we summed their areas.

We extracted adult body mass for each species from the PanTHERIA database[44], as larger species are often more vulnerable to anthropogenic impacts owing to their larger home ranges, slower life histories, and greater value to hunters[4,6]. We used body mass in lieu of specific attributes such as gestation length or weaning age because it represents a single integrated metric that is highly correlated with a wide range of life-history traits and ecological processes that have been linked with extinction proneness in mammals[6,7,45].

HPD has also been linked with mammalian extinction risk[6,7,24]. We extracted human population from of the Gridded Population of the World v.3 dataset provided in the PRIO-GRID socioeconomic variables table[46,47]. We divided numbers of people by grid-cell areas to obtain HPD and averaged across the grid cells overlapping each protected area over the time period corresponding to each λ estimate. These data were available only for 1990, 1995, 2000, and 2005. We therefore assigned population densities to other years using data from the nearest available year. For instance, HPD for the interval 1989–1995 was calculated as: $(4 \times \mathrm{HPD}_{1990}$ (for each year 1989–1992) $+ 3 \times \mathrm{HPD}_{1995}$ (for each year 1993–1995))/7 years.

We included Transparency International's Corruption Perceptions Index (CPI) to represent country-level governance quality[48], which has been shown to correlate with trends in elephant and rhinoceros populations[25]. Because CPI scores were available only for 1995–2010, we followed previous authors[25] in regressing annual CPI scores for each country × year combination against the mean of three International Country Risk Guide[49] variables (corruption, bureaucratic quality, and law-and-order) used to assess financial risk from government instability. We inferred missing CPI values on the basis of this regression ($R^2 = 0.80$) and averaged across the years spanned by each λ estimate. For three countries (Benin, Central African Republic, Chad) that had no scores for the International Country Risk Guide variables, we used CPI scores from the closest years available.

We included two predictors that captured the degree of urbanization in the vicinity of focal protected areas, based on the hypothesis that protected areas close to cities, such as Kenya's Nairobi National Park, might be more visited, better funded, and therefore better protected. First, we included percentage of urban area (derived from ref. 50) by averaging its values across the grid cells for each λ estimate. These data are a snapshot estimated for the year 2009 only, but should nonetheless capture broad patterns in urbanization dating back decades. Second, because the nearest urban areas were frequently located outside focal grid cells, we included travel time in minutes to the nearest urban centre of $\geq 50{,}000$ people (data from ref. 51); this is also a time-invariant predictor with values modelled for the years 1990–2005.

Drought is the main climatic driver of African mammal population trends[52–54]. We therefore included an annualized version of the Standardized Precipitation and Evapotranspiration Index (SPEI) as a metric of drought frequency[26,47]. We calculated the proportion of months in each year with a SPEI value ≥ 1.5 s.d. below the long-term mean to measure the frequency of drought in each year[47,55] and averaged this value across all grid-cell-years for each λ estimate.

Last, because mining has been shown to negatively affect local biodiversity[17], we created a binary variable indicating the presence or absence of extractable mineral resources. This was coded as '1' for parks overlapping grid cells with known deposits of gold, diamonds, petroleum, or precious gemstones. Data were drawn from PRIO-GRID v.2.0[47].

Prior to analysis, we tested the correlation between each pair of predictors; we considered $r > 0.70$ to represent a potentially problematic level of collinearity[56]. All pairwise combinations had $r < 0.43$.

Statistical modelling. Analyses were conducted using R v.3.0[57]. We used linear regression together with information-theoretic model-selection and model-averaging approaches, based on AIC$_c$, to identify the most appropriate explanatory models for λ and to calculate robust estimates of model parameters[22,58,59]. AIC$_c$, which measures the quality of a given model relative to the others in a candidate set, balances goodness-of-fit against complexity by penalizing the inclusion of each additional parameter; lower values of AIC$_c$ signify better models[22]. Akaike weights, w_i, give the probability that a model is the best in the candidate set and can be used to calculate RVI[22]. Multi-model averaging enables more robust inferences by accounting for uncertainty as to which is the true best model: a single set of parameter estimates is obtained by taking the weighted mean of the parameter estimates for each variable across the models retained in the 95% confidence set, with the contribution of each model weighted by its w_i (refs 22, 58).

Prior to model-averaging, the predictor and response variables were centred (by subtracting the mean from each value) and standardized (by dividing the centred value by the partial standard deviation)[58] using the R package MuMIn[60]. This approach rescales the variables so that their means are 0 and standard deviations are 1 and expresses all parameter estimates as the rate of change in λ per s.d. change in the predictor, which facilitates comparison of β values and is necessary for the accurate calculation of model-averaged parameter estimates[58]. This approach also causes the regression line to pass through the origin (see Table 1), because the value of the fitted standardized λ is at its mean (0) when all of the predictors are at their means ($x = 0$).

We report model AIC$_c$, ΔAIC$_c$, R^2, and w_i values (Extended Data Tables 3, 4). For individual predictors, we report standardized model-averaged parameter estimates, standard errors, and 95% confidence intervals, which indicate statistical significance ($P < 0.05$) when not overlapping zero[60] (Table 1).

Model validation. We validated the top regression model for each interval in light of the assumptions of linear regression by inspecting diagnostic plots of the model residuals (Extended Data Figs 2, 3). We used histograms of residuals to evaluate their normality, plots of the residuals against fitted λ values to check for homogeneity of variance, and plots of standardized residuals against leverage to identify data points with disproportionate influence on the regression results[61]. We found no indication that these assumptions were violated, except for the presence of one possible outlier in the 1989–2010 dataset (Extended Data Fig. 3c). Removing this data point had no appreciable impact on our results. Also, we plotted regression residuals against the value of each predictor variable in the top models; patterned residuals in these plots would have indicated the need for non-additive (interaction) or nonlinear terms in our models, but no such patterns were observed[61].

For graphical representation of the effect of conflict frequency on λ, we used the R package visreg[62] to generate conditional-regression plots (Fig. 2), which estimate the relationship between conflict frequency and λ in the best-fitting model for each interval while holding other predictors (body mass and protected area size for 1946–2010; HPD, drought, and percentage of urban area for 1989–2010) constant at their medians.

Sensitivity analyses. We tested the sensitivity of our analyses to several aspects of our data and analytical approach. First, because elephants were disproportionately represented ($n = 97$ of 253 records for 1946–2010; $n = 94$ of 172 records for 1989–2010), we tested the robustness of our conclusions when elephants were excluded from the dataset. For the full time period, conflict frequency remained a significant (and slightly stronger) negative predictor of λ (the model-averaged standardized parameter estimate decreased from -0.06 to -0.09). For the restricted period, the model-averaged standardized conflict-frequency parameter remained negative (-0.07) albeit only marginally statistically significant (confidence interval: -0.17, 0.001), probably because statistical power was lost by halving the dataset. However, the single best-fitting model still included a highly significant effect of conflict frequency ($\beta_{\mathrm{CF}} = -0.11$, $P < 0.0001$) and explained considerably more variance than did the best-fitting model with elephants included ($R^2 = 0.38$, vs. 0.10). We therefore believe that our conclusions are not altered by the preponderance of elephant records.

Second, we tested the robustness of our results to the use of an alternative model-selection procedure (Supplementary Table 3), namely general-to-specific modelling[63,64], an automated stepwise predictor-addition and -deletion algorithm. For both time periods, the same predictors were retained as in our original analysis, with very similar parameter estimates.

Third, we used a bootstrapping method to assess the sensitivity of our findings to the decision rules used for selecting a single λ estimate in cases where the data included more than one such estimate for a given population. We re-computed the regression-parameter estimates for the best-fitting model in each time period using 10,000 bootstrap replicates, each time randomly selecting a single λ value from those available for each population. We compared the bootstrapped 95% confidence intervals for these regression-parameter estimates with those from our original analyses. For all significant predictors (conflict frequency for 1946–2010, conflict frequency and HPD for 1989–2010), the bootstrapped intervals encompassed the original parameter estimates, indicating that our conclusions were robust to the duplicate-elimination procedure (Extended Data Table 2).

Fourth, we tested for residual spatial autocorrelation in the best-fitting model for both time periods. We calculated a standard metric of spatial autocorrelation, Moran's I (refs 65, 66), for all pairwise combinations of data points from the model residuals and plotted I as a function of the geographical distance between protected areas. $I = 0$ indicates a random distribution of values in space, $I > 0$ indicates spatial clustering, and $I < 0$ indicates overdispersion. The magnitude of I can be interpreted similarly to a correlation coefficient, with $I < 0.20$ considered to indicate that little autocorrelation is present[66]. If autocorrelation is present (that is, if there is an unaccounted-for spatially correlated process influencing the data), we would expect a monotonically decreasing trend in I as the distance between sampled locations increases[65,67]. For both the full- and restricted-interval datasets, we found no such evidence of residual spatial autocorrelation: the magnitude of I was always below 0.20, and I exhibited no directional trend with increasing distance between locations (Extended Data Fig. 4).

Fifth, although we found no evidence for marked residual spatial autocorrelation at any distance (including 0 km, where more than one data point came from the same protected area), we conducted an additional independent check on the robustness of our results to potential pseudoreplication arising from inclusion of more than one species from certain protected areas. We used a bootstrapping technique (nearly identical to that described above) to iteratively recalculate parameter estimates 10,000 times for the top regression model in each time period, randomly selecting only one species per protected area in each iteration (or two in the few cases where the date ranges of the λ estimates were entirely non-overlapping), which prevented potentially non-independent records from being included together in the same analysis. We found that the effect of conflict frequency was largely robust to this procedure, suggesting that our central result was not an artefact of pseudoreplication (Extended Data Table 5).

Sixth, we tested the robustness of our results to a coarsening of the conflict data, to account for possible inaccuracies in the location of geocoded conflicts in the PRIO-GRID and GED products. We achieved this by devising a national-level conflict-frequency metric (CF$_{\mathrm{national}}$), estimated as the proportion of years between each λ estimate's starting and ending years in which at least one PRIO-GRID/GED conflict occurred anywhere in the country. The significant negative model-averaged effect of conflict was maintained for the full time period (Extended Data Table 6). For the restricted period, the 95% confidence interval of the model-averaged CF$_{\mathrm{national}}$ parameter slightly overlapped zero, implying marginal statistical significance, and the CF$_{\mathrm{national}}$ effect was also marginally significant in the single top model (unstandardized $\beta = -0.12$, $P = 0.06$). We interpret these results as indicating that our original, local conflict frequency metric was the more consistent predictor of λ, but that coarsening the conflict data to the national scale nonetheless corroborates our central inference about the negative effect of conflict frequency on λ.

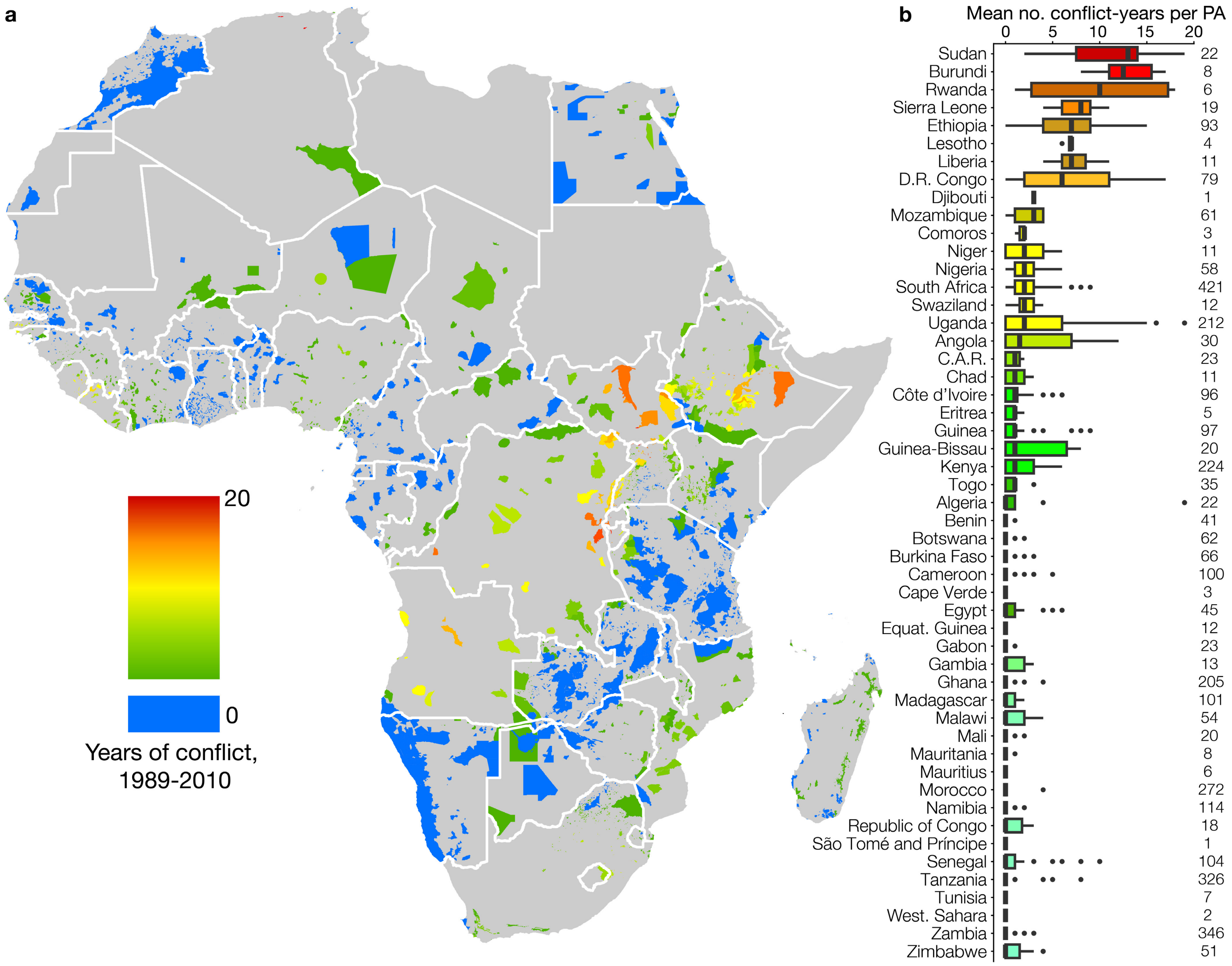
Seventh, we tested two alternative formulations of conflict frequency that isolated the spatial (maximum proportion of grid cells in conflict in any single year) and temporal (proportion of years with one or more grid cell in conflict) dimensions of the original conflict-frequency metric. For both time periods, the two new metrics were highly correlated with each other ($r \geq 0.84$) and with our original conflict-frequency metric ($r \geq 0.89$), making it inappropriate to include both metrics within a single candidate-model set[58]. However, pairwise correlations with λ were stronger for CF$_{\mathrm{spatial}}$ than for CF$_{\mathrm{temporal}}$ ($r_{\text{spatial-1946–2010}} = -0.25$, $P < 0.001$; $r_{\text{temporal-1946–2010}} = -0.20$, $P = 0.001$; $r_{\text{spatial-1989–2010}} = -0.20$, $P = 0.01$; $r_{\text{temporal-1989–2010}} = -0.10$, $P = 0.17$). In the future, higher-resolution conflict data may allow more precise decomposition of the spatial and temporal dimensions of conflict.

Eighth, we investigated the spatial scale over which conflict affects large-mammal populations by calculating spatially lagged conflict frequency over four nested spatial scales. Lags were introduced by calculating the conflict frequency for grid cells overlapping the focal protected areas plus that of their surrounding neighbourhoods extending one, two, three, or four grid cells in all directions. At all four of these spatial scales, lagged conflict frequencies were extremely highly correlated with the original conflict frequency ($r > 0.95$) in our dataset. Thus, changing the spatial scale over which we assess conflict impacts on λ is unlikely to change our results.

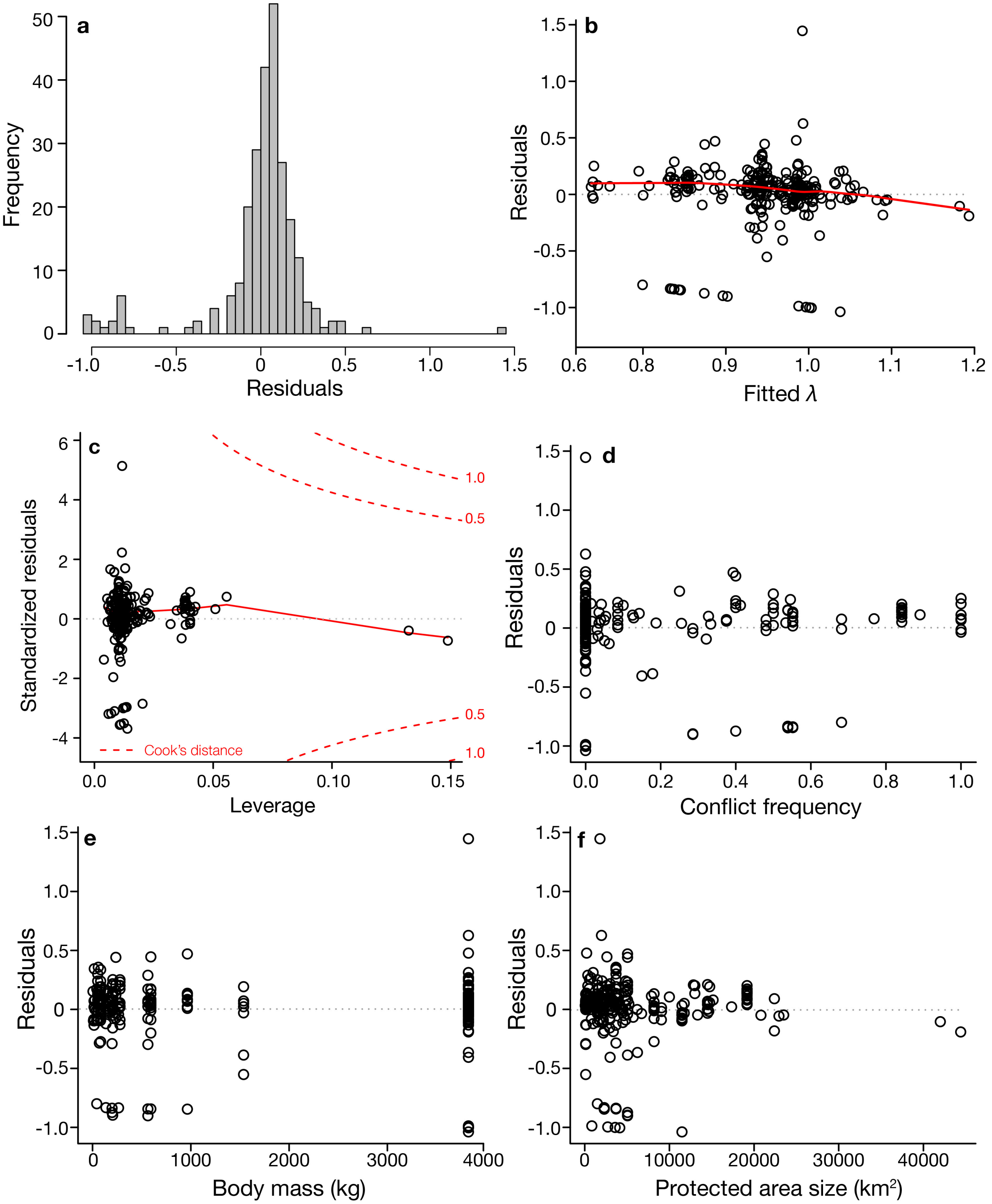
Last, we checked for delayed effects of drought on λ by computing temporally lagged drought-frequency variables for intervals starting 1, 2, 5, 10, 15, and 20 years before the interval associated with each λ. These lagged drought terms were generally highly correlated with the original (un-lagged) drought-frequency metric, and substituting lagged drought terms for the original in the best-fitting model for 1989–2010 did not improve model fit.

Data availability. The data that support the findings of this study are provided, along with the original sources from which raw wildlife-population data were extracted, in Supplementary Tables 4 and 5 and Supplementary Data 1 and 2. All raw data for protected areas and other predictor variables were obtained from publicly available sources[18,20,21,44,46–48]. Replicating the bootstrap sensitivity analyses in Extended Data Tables 2 and 5 requires the larger wildlife-population database compiled by the lead author, which is available upon request.

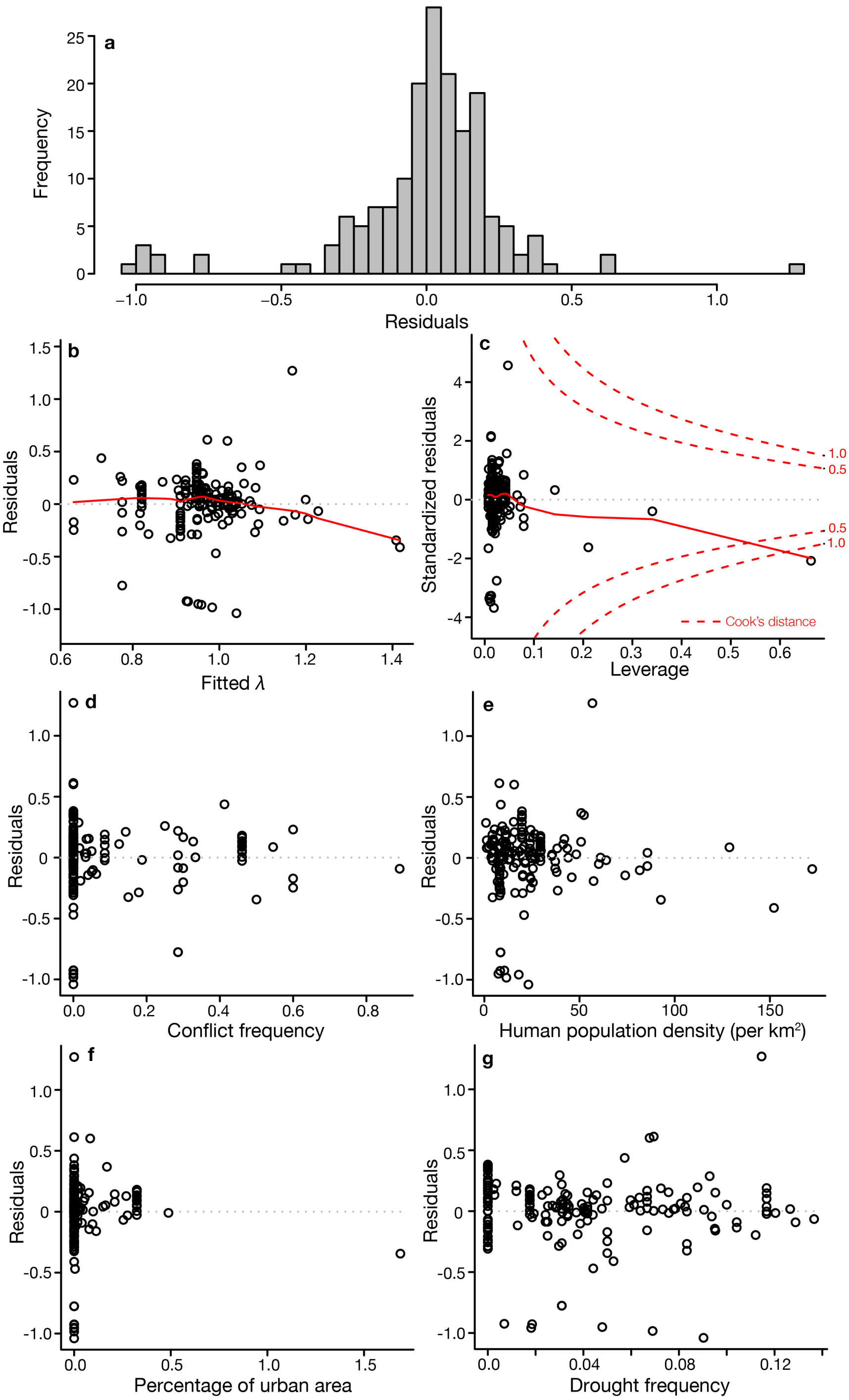
31. Wittemyer, G. *et al.* Illegal killing for ivory drives global decline in African elephants. *Proc. Natl Acad. Sci. USA* **111,** 13117–13121 (2014).
32. Blanc, J. J. *et al.* Changes in elephant numbers in major savanna populations in eastern and southern Africa. *Pachyderm* **38,** 19–28 (2005).
33. Croicu, M. C. & Sundberg, R. *UCDP GED Conflict Polygons Dataset Codebook Version 1.1-2011* (Department of Peace and Conflict Research, Uppsala Univ., 2012).
34. von Uexkull, N., Croicu, M., Fjelde, H. & Buhaug, H. Civil conflict sensitivity to growing-season drought. *Proc. Natl Acad. Sci. USA* **113,** 12391–12396 (2016).
35. Schleussner, C.-F., Donges, J. F., Donner, R. V. & Schellnhuber, H. J. Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries. *Proc. Natl Acad. Sci. USA* **113,** 9216–9221 (2016).
36. Esteban, J., Mayoral, L. & Ray, D. Ethnicity and conflict: theory and facts. *Science* **336,** 858–865 (2012).
37. Burke, M. B., Miguel, E., Satyanath, S., Dykema, J. A. & Lobell, D. B. Warming increases the risk of civil war in Africa. *Proc. Natl Acad. Sci. USA* **106,** 20670–20674 (2009).
38. Kanyamibwa, S. Impact of war on conservation: Rwandan environment and wildlife in agony. *Biodivers. Conserv.* **7,** 1399–1406 (1998).
39. Baldus, R. D., Hahn, R., Ellis, C. & DeLeon, S. D. in *Peace Parks: Conservation and Conflict Resolution* (ed. Ali, S. H.) 109–126 (MIT Press, 2007).
40. Bouché, P. *et al.* Game over! Wildlife collapse in northern Central African Republic. *Environ. Monit. Assess.* **184,** 7001–7011 (2012).
41. Raleigh, C., Linke, A., Hegre, H. & Karlsen, J. Introducing ACLED: an armed conflict location and event dataset special data feature. *J. Peace Res.* **47,** 651–660 (2010).
42. Eck, K. In data we trust? A comparison of UCDP GED and ACLED conflict events datasets. *Coop. Confl.* **47,** 124–141 (2012).
43. Woodroffe, R. & Ginsberg, J. R. Edge effects and the extinction of populations inside protected areas. *Science* **280,** 2126–2128 (1998).
44. Jones, K. E. *et al.* PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90,** 2648 (2009).
45. Brashares, J. S. Ecological, behavioral, and life-history correlates of mammal extinctions in West Africa. *Conserv. Biol.* **17,** 733–743 (2003).
46. Center for International Earth Science Information Network & International Center for Tropical Agriculture. Gridded population of the world, v.3. https://dx.doi.org/10.7927/H4XK8CG2 (NASA Socioeconomic Data and Applications Center, 2005).
47. Tollefsen, A. F., Bahgat, K., Nordkvelle, J. & Buhaug, H. *PRIO-GRID codebook v.2.0* (Peace Research Institute Oslo, 2015).
48. Transparency International. Corruption Perceptions Index. https://www.transparency.org/news/feature/corruption_perceptions_index_2016 (2016).
49. P.R.S. Group. International Country Risk Guide. http://epub.prsgroup.com/products/international-country-risk-guide-icrg (2006).
50. Bontemps, S. *et al. GLOBCOVER 2009-Products Description and Validation Report* (Univ. Catholique de Louvain and European Space Agency, 2011).
51. Uchida, H. & Nelson, A. *Agglomeration Index: Towards a New Measure of Urban Concentration* (United Nations University World Institute for Development Economics Research, 2010).
52. Ogutu, J. O. & Owen-Smith, N. ENSO, rainfall and temperature influences on extreme population declines among African savanna ungulates. *Ecol. Lett.* **6,** 412–419 (2003).
53. Ogutu, J. O., Piepho, H.-P., Dublin, H. T., Bhola, N. & Reid, R. S. Rainfall influences on ungulate population abundance in the Mara–Serengeti ecosystem. *J. Anim. Ecol.* **77,** 814–829 (2008).
54. Augustine, D. J. Response of native ungulates to drought in semi-arid Kenyan rangeland. *Afr. J. Ecol.* **48,** 1009–1020 (2010).
55. Beguería, S., Vicente-Serrano, S. M., Reig, F. & Latorre, B. Standardized precipitation evapotranspiration index (SPEI) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *Int. J. Climatol.* **34,** 3001–3023 (2014).
56. Dormann, C. F. *et al.* Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36,** 27–46 (2013).
57. R Core Team. *R: a Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2013).
58. Cade, B. S. Model averaging and muddled multimodel inferences. *Ecology* **96,** 2370–2382 (2015).
59. Grueber, C. E., Nakagawa, S., Laws, R. J. & Jamieson, I. G. Multimodel inference in ecology and evolution: challenges and solutions. *J. Evol. Biol.* **24,** 699–711 (2011).
60. Bartoń, K. Package 'MuMIn': Multi-model Inference. R Package v.1.15.6 https://cran.r-project.org/web/packages/MuMIn/index.html (2015).
61. Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A. & Smith, G. M. *Mixed Effects Models and Extensions in Ecology with R* (Springer, 2009).
62. Breheny, P. & Burchett, W. visreg: Visualization of regression models. R Package v.2.0 https://cran.r-project.org/web/packages/visreg/index.html (2012).
63. Campos, J., Ericsson, N. R. & Hendry, D. F. *General-to-Specific Modeling: an Overview and Selected Bibliography* (Board of Governors of the Federal Reserve System, 2005).
64. Sucarrat, G., Pretis, F. & Reade, J. gets: General-to-Specific (GETS) modelling and indicator saturation methods. R package v.0.1. https://CRAN.R-project.org/package=gets (2017).
65. Lichstein, J. W., Simons, T. R., Shriner, S. A. & Franzreb, K. E. Spatial autocorrelation and autoregressive models in ecology. *Ecol. Monogr.* **72,** 445–463 (2002).
66. Turner, M. G., Gardner, R. H. & O'Neill, R. V. *Landscape Ecology in Theory and Practice: Pattern and Process* (Springer, 2015).
67. Diniz-Filho, J. A. F., Bini, L. M. & Hawkins, B. A. Spatial autocorrelation and red herrings in geographical ecology. *Glob. Ecol. Biogeogr.* **12,** 53–64 (2003).
68. Draper, N. R & Smith, H. in *Applied Regression Analysis* Ch. 2 (Wiley-Interscience, 1998).
69. Kühn, I. & Dormann, C. F. Less than eight (and a half) misconceptions of spatial analysis. *J. Biogeogr.* **39,** 995–998 (2012).



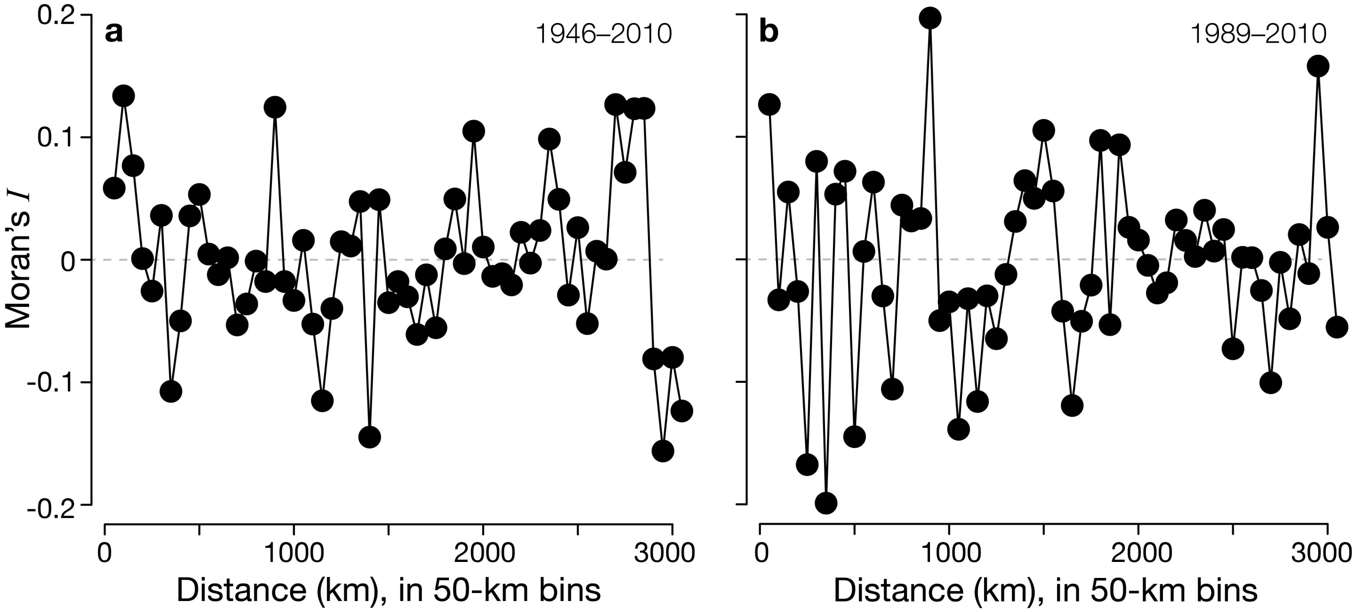
Extended Data Figure 1 | Distribution and frequency of armed conflict in African protected areas for the restricted interval, 1989–2010. a, Number of conflict-years in each protected area; colours indicate average value across all grid cells overlapping the protected area. **b**, Mean conflict-years per protected area in each country. Boxes, inter-quartile ranges; vertical lines, medians; whiskers, 1.5× the inter-quartile range from the median; dots, outlying values. Total number of protected areas included per country, from the World Database of Protected Areas[21], is shown on the right; statistical analyses of the correlation between conflict and wildlife population trajectories were conducted using the subset of these protected areas for which adequate wildlife data were obtainable. Sudan and South Sudan are distinguished in **a** but combined in **b**; two outlying island nations, Cape Verde and Mauritius, are omitted from **a** but included in **b**. Map created in ArcGIS and R using open-access country-border data from the Global Administrative Areas database (https://gadm.org). C.A.R., Central African Republic; D.R. Congo, Democratic Republic of Congo; Equat. Guinea, Equatorial Guinea; West. Sahara, Western Sahara.



Extended Data Figure 2 | Regression validation plots for the top model for 1946–2010. Model: $\lambda \sim$ conflict frequency + body mass + protected-area size. **a**, Histogram showing approximate normality of regression residuals. **b**, Plot of residuals versus model fit, showing no clear pattern (thus, no pronounced heteroscedasticity). **c**, Standardized residuals versus leverage; dashed red lines show Cook's distance contours, indicating that no points exerted disproportionate influence on the regression outcome (that is, no points had Cook's distance > 1.0)[61]. **d**–**f**, Plots of residuals versus conflict frequency (**d**), body mass (**e**), and protected-area size (**f**). All show no clear pattern, which validates our decision not to include non-additive interaction terms in the candidate-model set[61]. They also show no curvilinear relationships with the predictor, which indicates that there is no strong justification for including nonlinear fits[68]. **a**–**f**, Data are from 253 λ estimates.



Extended Data Figure 3 | Regression-validation plots for the top model for 1989–2010. Model: $\lambda \sim$ conflict frequency + HPD + percentage of urban area + drought frequency. **a**, Histogram showing approximate normality of regression residuals. **b**, Plot of residuals versus model fit, showing no clear pattern (thus, no pronounced heteroscedasticity). **c**, Standardized residuals versus leverage; dashed red lines show Cook's distance contours, indicating that just one point may have exerted disproportionate influence on the regression outcome (Cook's distance >1.0)[61]. However, excluding this datum did not qualitatively alter the results. **d**–**f**, Plots of residuals versus conflict frequency (**d**), HPD (**e**), percentage of urban area (**f**), and SPEI drought index (**g**). All show no clear pattern, which validates our decision not to include non-additive interaction terms in the candidate-model set[61], and also show no curvilinear relationships, which indicates no strong justification for including nonlinear fits[68]. **a**–**f**, Data are from 172 λ estimates.



Extended Data Figure 4 | Moran's *I* plots testing for residual spatial autocorrelation. **a**, **b**, Moran's *I* (ref. 65) for all pairwise combinations of data points, calculated from the residuals of the best-fitting models[69] for 1946–2010 (**a**) and 1989–2010 (**b**) and plotted as a function of the geographical distance between the protected areas from which the data were drawn (in 50-km bins). $I = 0$ indicates a random distribution of values in space, $I > 0$ indicates spatial clustering, and $I < 0$ indicates overdispersion. The magnitude of *I* can be interpreted similarly to a correlation coefficient, with $I < 0.20$ considered to indicate that little autocorrelation is present at a given distance class[66]. The absence of a monotonically decreasing trend in *I* as the distance between sampled locations increases supports our interpretation that λ did not co-vary as a function of some unaccounted-for underlying spatial process that might confer statistical non-independence[67].

Extended Data Table 1 | Methodology for literature search

ISI Web of Science literature search terms (April 29, 2015)
ts=Park OR ts=protected area OR ts=conservation area OR ts=wildlife management area OR ts=World Heritage
AND
ts=“Population size” OR ts=“population number” OR ts=“density” OR ts=“encounter rate” OR ts=(aerial AND count OR sampl*) OR ts="game count" OR ts="individual registration"
AND
ts=Ungulate OR ts=antelope OR ts=*buck OR ts=*bok OR ts=hartebeest OR ts=giraffe OR ts=elephant OR ts=lion OR ts=buffalo OR ts=leopard OR ts=cheetah OR ts=zebra OR ts=primate OR ts=monkey OR ts=baboon OR ts=gorilla OR ts=pig OR ts=warthog OR ts=hyena OR ts=caracal OR OR ts=fox OR ts=rhinoceros OR ts=hippopotamus OR ts=eland OR ts=kudu OR ts=gazelle OR ts=topi OR ts=jackal OR ts=“wild dog”

This initial literature search was designed to yield an unbiased set of publications reporting population densities for large African wildlife; our initial search included terms pertinent to primates and carnivores, but we found that these species were often surveyed using divergent methodologies (often in ways that precluded density estimation) or did not meet our quality-control criteria.
‘ts’ denotes a ‘topic search’ in Web of Science, which searches titles, abstracts, and keywords. Boolean operators, parentheses, and truncation (*) were used to refine the search set. We subsequently reviewed the references yielded by this initial search to identify additional relevant sources, databases, and grey literature (see Methods: Obtaining population-density estimates from the literature).

Extended Data Table 2 | Bootstrap analysis of the sensitivity of the results of the best-fitting model for each interval to the sequential duplicate-λ filtering process

	Original top model		**Bootstrap analysis**	
1946–2010	β	**s.e.**	β_{LCL}	β_{UCL}
Intercept	0.93	0.03	0.93	0.99
Conflict frequency*	–0.20	0.06	–0.34	–0.17
Protected-area size	4.7×10^{-6}	2.7×10^{-6}	1.1×10^{-6}	5.4×10^{-6}
Body mass	1.5×10^{-5}	1.1×10^{-5}	2.9×10^{-6}	1.8×10^{-5}
1989–2010	β	**s.e.**	β_{LCL}	β_{UCL}
Intercept	0.88	0.04	0.88	0.96
Conflict frequency*	–0.57	0.14	–0.75	–0.37
HPD*	0.0032	0.00096	0.0019	0.0034
Percentage of urban area	0.28	0.15	–0.07	0.19
Drought frequency	0.90	0.61	–0.03	1.16

For all significant predictors in our primary model-averaged analyses (indicated by asterisks), the bootstrapped lower (β_{LCL}) and upper (β_{UCL}) 95% confidence limits around the coefficients from the single best-fitting model for each interval (obtained using randomly selected λ estimates for each population) encompassed the original parameter estimates (β) (Extended Data Tables 3, 4). Thus, our results were not qualitatively affected by the filtering procedure used to select among duplicate λ estimates (see Methods: Sensitivity analyses). HPD, human population density. Data are from 253 λ estimates for 1946–2010 and 172 λ estimates for 1989–2010.

Extended Data Table 3 | Full candidate-model set and model-selection criteria, 1946–2010

Rank	Model specification	AIC$_c$	ΔAIC$_c$	w_i	R^2
1*	CF + BM + PA size	85.64	0.00	0.31	0.05
2*	CF + PA size	85.65	0.01	0.31	0.05
3*	CF + BM	86.64	1.00	0.19	0.05
4*	CF	86.71	1.08	0.18	0.04
5	BM	92.94	7.30	0.01	0.02
6	PA Size	93.04	7.40	0.01	0.00
7	NULL	95.52	9.89	0.00	0.00
8	PA Size	95.71	10.01	0.00	0.01

The top four models (indicated by asterisks next to rank) represent the 95% confidence set used in model averaging (Table 1). BM, body mass; CF, conflict frequency; NULL, intercept-only model; PA size, protected-area size. Data are from 253 λ estimates.

Extended Data Table 4 | Top models and model-selection criteria, 1989–2010

Rank	Model specification	AIC$_c$	ΔAIC$_c$	w_i	R^2
1	CF + HPD + Drought + Urban	63.50	0.00	0.03	0.13
2	CF + HPD + Urban	63.60	0.10	0.02	0.11
3	CF + HPD + Urban + BM	63.92	0.41	0.02	0.12
4	CF + HPD + Drought + Urban + Resource	64.01	0.51	0.02	0.13
5	CF + HPD + BM	64.20	0.70	0.02	0.11
6	CF + HPD	64.26	0.75	0.02	0.10
7	CF + HPD + Urban + Resource	64.55	1.05	0.02	0.12
8	CF + HPD + Drought	64.73	1.23	0.01	0.11
9	CF + HPD + Urban + BM + Resource	64.85	1.35	0.01	0.13
10	CF + HPD + Urban + BM	64.89	1.39	0.01	0.13
11	CF + HPD + Drought + Urban + PA size + Resource	64.99	1.49	0.01	0.14
12	CF + HPD + BM + Resource	65.04	1.54	0.01	0.12
13	CF + HPD + Resource	65.11	1.61	0.01	0.11
14	CF + HPD + Drought + Resource	65.20	1.70	0.01	0.12
15	CF + HPD + Drought + Urban + PA size	65.31	1.80	0.01	0.13
16	CF + HPD + Drought + Urban + CPI	65.42	1.92	0.01	0.13
17	CF + HPD + Drought + Urban + BM + Resource	65.50	2.00	0.01	0.14

Shown are the top 17 models from the full set ($n = 1{,}024$ models, based on 172 λ estimates), including all those within $\Delta\text{AIC}_c \le 2.0$ of the single best-fitting model. All models shown here were included in the 95% confidence set (Table 1), which comprised 240 total models. BM, body mass; CF, conflict frequency; CPI, Corruption Perceptions Index; Drought, drought frequency; HPD, human population density; PA size, protected-area size; Resource, presence of extractable mineral resources; Urban, percentage of urban area.

Extended Data Table 5 | Bootstrap analysis of the sensitivity of the results of the best-fitting model for each interval to the inclusion of λ values for multiple species from the same protected area

	Original top model		Bootstrap analysis	
1946–2010	β	**s.e.**	β_{LCL}	β_{UCL}
Intercept	0.93	0.03	0.86	0.99
Conflict frequency*	–0.20	0.06	–0.29	0.02
Protected-area size	4.7×10^{-6}	2.8×10^{-6}	8.2×10^{-7}	2.8×10^{-6}
Body mass	1.5×10^{-5}	1.1×10^{-5}	2.8×10^{-6}	3.7×10^{-5}
1989–2010	β	**s.e.**	β_{LCL}	β_{UCL}
Intercept	0.88	0.04	0.90	0.99
Conflict frequency*	–0.57	0.14	–0.59	–0.22
HPD*	0.0032	0.00096	–0.16	0.84
Percentage of urban area	0.28	0.15	0.07	0.16
Drought frequency	0.90	0.61	0.0017	0.0029

For all terms in the best-fitting model for each interval (Extended Data Tables 3, 4), we show the original parameter estimates (β) and their standard errors, along with bootstrapped lower (β_{LCL}) and upper (β_{UCL}) 95% confidence limits obtained when only one or two randomly selected λ estimates were included for any given protected area (see Methods: Sensitivity analyses). Predictors indicated by asterisks were significant in the primary model-averaged analyses reported in the main text (Table 1). Data are from 105 protected areas or protected-area complexes for 1946–2010 and 96 protected areas or protected-area complexes for 1989–2010.

Extended Data Table 6 | Model-averaged parameter estimates using the spatially coarsened, national-level conflict frequency ($CF_{national}$)

1946–2010	β	**s.e.**	β_{LCL}	β_{UCL}	**RVI**
Intercept	0.00	0.00	N/A	N/A	N/A
$CF_{national}$*	–0.05	0.02	–0.09	–0.02	0.97
Protected-area size	0.01	0.02	–0.01	0.06	0.76
Body mass*	0.03	0.02	0.002	0.07	0.50
1989–2010	β	**s.e.**	β_{LCL}	β_{UCL}	**RVI**
Intercept	0.00	0.00	N/A	N/A	N/A
Body mass*	0.04	0.03	0.01	0.10	0.80
HPD*	0.04	0.03	0.004	0.10	0.78
$CF_{national}$	–0.03	0.03	–0.09	0.01	0.62
Resource presence	–0.01	0.02	–0.08	0.02	0.44
Travel time to urban area	–0.01	0.02	–0.08	0.02	0.42
Drought frequency	0.01	0.02	–0.02	0.07	0.40
Conflict intensity	–0.01	0.02	–0.08	0.03	0.34
Percentage of urban area	0.01	0.02	–0.03	0.06	0.32
Protected-area size	0.003	0.01	–0.04	0.06	0.29
Corruption index	–0.0002	0.01	–0.05	0.05	0.26

Standardized and centred[58] model-averaged parameter estimates (β), standard errors, upper (β_{LCL}) and lower (β_{UCL}) 95% confidence limits, and RVI for each predictor in 1946–2010 (253 λ estimates) and 1989–2010 (172 λ estimates). Predictors are listed in descending order of RVI within each interval; those that were statistically significant in this analysis (that is, where β_{LCL} and β_{UCL} did not overlap 0) are indicated by asterisks.

doi:10.1038/nature25181

A global map of travel time to cities to assess inequalities in accessibility in 2015

D. J. Weiss[1], A. Nelson[2], H. S. Gibson[1], W. Temperley[3], S. Peedell[3], A. Lieber[4], M. Hancher[4], E. Poyart[4], S. Belchior[5], N. Fullman[6], B. Mappin[7], U. Dalrymple[1], J. Rozier[1], T. C. D. Lucas[1], R. E. Howes[1], L. S. Tusting[1], S. Y. Kang[1], E. Cameron[1], D. Bisanzio[1], K. E. Battle[1], S. Bhatt[8] & P. W. Gething[1]

The economic and man-made resources that sustain human wellbeing are not distributed evenly across the world, but are instead heavily concentrated in cities. Poor access to opportunities and services offered by urban centres (a function of distance, transport infrastructure, and the spatial distribution of cities) is a major barrier to improved livelihoods and overall development. Advancing accessibility worldwide underpins the equity agenda of 'leaving no one behind' established by the Sustainable Development Goals of the United Nations[1]. This has renewed international efforts to accurately measure accessibility and generate a metric that can inform the design and implementation of development policies. The only previous attempt to reliably map accessibility worldwide, which was published nearly a decade ago[2], predated the baseline for the Sustainable Development Goals and excluded the recent expansion in infrastructure networks, particularly in lower-resource settings. In parallel, new data sources provided by Open Street Map and Google now capture transportation networks with unprecedented detail and precision. Here we develop and validate a map that quantifies travel time to cities for 2015 at a spatial resolution of approximately one by one kilometre by integrating ten global-scale surfaces that characterize factors affecting human movement rates and 13,840 high-density urban centres within an established geospatial-modelling framework. Our results highlight disparities in accessibility relative to wealth as 50.9% of individuals living in low-income settings (concentrated in sub-Saharan Africa) reside within an hour of a city compared to 90.7% of individuals in high-income settings. By further triangulating this map against socioeconomic datasets, we demonstrate how access to urban centres stratifies the economic, educational, and health status of humanity.

An axiom for the twenty-first century is that the world is becoming increasingly connected. Although this is certainly true for electronic forms of communication, physical links between locations, and thus the time it takes to move between them, remain constrained by available infrastructure as well as physical and political impediments to travel. Eliminating disparities in accessibility is central to the Sustainable Development Goals (SDGs) set out by the United Nations[1], which explicitly call for improved or universal access to key services, including education programs, health services, and banking and financial institutions. Cities are the epicentres of such activities[3–6], and how easily people can reach urban areas directly affects whether crucial services can be obtained.

What constitutes accessibility is widely debated and precise definitions of this metric can be arbitrary. In this study, we operationalize accessibility in terms of travel time required to reach the nearest urban centre, defined as a contiguous area with 1,500 or more inhabitants per square kilometer or a majority of built-up land cover coincident with a population centre of at least 50,000 inhabitants[7]. We define accessibility using travel time as it is readily interpretable, can feasibly be generated at global scales, and is known to be a predictive metric in research domains including conservation[8], food security[6], trade[9] and population health[10,11]. Furthermore, travel time better captures the opportunity cost of travel than Euclidean or network distance, and ultimately reflects the information humans use to inform transport decisions. The outcome of this research is a map that provides an actionable dataset that will support many research and policy needs. To demonstrate the map's utility for global and local decision-making, we provide exploratory analyses examining relationships between accessibility and national-level income as well as economic prosperity, educational attainment, and healthcare utilization at the level of household clusters.

Our study responds to an increased need for fine-grained quantification of accessibility worldwide. The only previous assessment of global accessibility[2] was for the year 2000, and marked advances in data quality and availability have since occurred. By anchoring our global accessibility map to 2015 (that is, the year of formal SDG adoption), we provide a baseline from which to track infrastructural improvements and urban expansion throughout the duration of the SDGs because accessibility is a precondition for many development targets. Although our results are useful in a variety of contexts, their potential impact centres around a more unifying aim: catalysing action to narrow gaps in opportunity by improving accessibility for remote populations and/or reducing disparities between populations with differing degrees of connectivity to cities.

To quantify travel time required to reach the nearest city via surface transport (air travel is not considered), we applied a similar methodology to that which had been used previously to produce the foremost existing global accessibility map[2] to updated and expanded input data sets for 2015. These inputs consisted of gridded surfaces that quantify the geographical positions and salient attributes of roads, railways, rivers, water bodies, land cover types, topographical conditions (slope angle and elevation), and national borders. Roads are the primary driver of accessibility globally and also represent the most substantial advance from the previous accessibility mapping effort. Our roads dataset was created by merging Open Street Map (OSM) data with a distance-to-roads product derived from the Google roads database; these datasets were extracted in November 2016 and March 2016, respectively. The resulting roads dataset represents a global-scale synthesis of these two data sources, and the unparalleled data quality of this dataset led to a 4.8-fold increase in road pixels relative to the dataset used for the previous accessibility map[2]. Although roads built since 2000 contributed to the increased data volume, the primary driver of this increase

[1]Malaria Atlas Project, Big Data Institute, Nuffield Department of Medicine, University of Oxford, Roosevelt Drive, Oxford OX3 7FY, UK. [2]Department of Natural Resources, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. [3]European Commission, Joint Research Centre, Unit D6 Knowledge for Sustainable Development and Food Security, Via Fermi 2749, Ispra 21027, Varese, Italy. [4]Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA. [5]Vizzuality, Office D, Dales Brewery, Gwydir Street, Cambridge CB1 2LJ, UK. [6]Institute for Health Metrics and Evaluation, University of Washington, 2301 5th Avenue, Suite 600, Seattle, Washington 98121, USA. [7]Centre for Biodiversity and Conservation Science, School of Biological Sciences, University of Queensland, St Lucia, Queensland 4072, Australia. [8]Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG, UK.

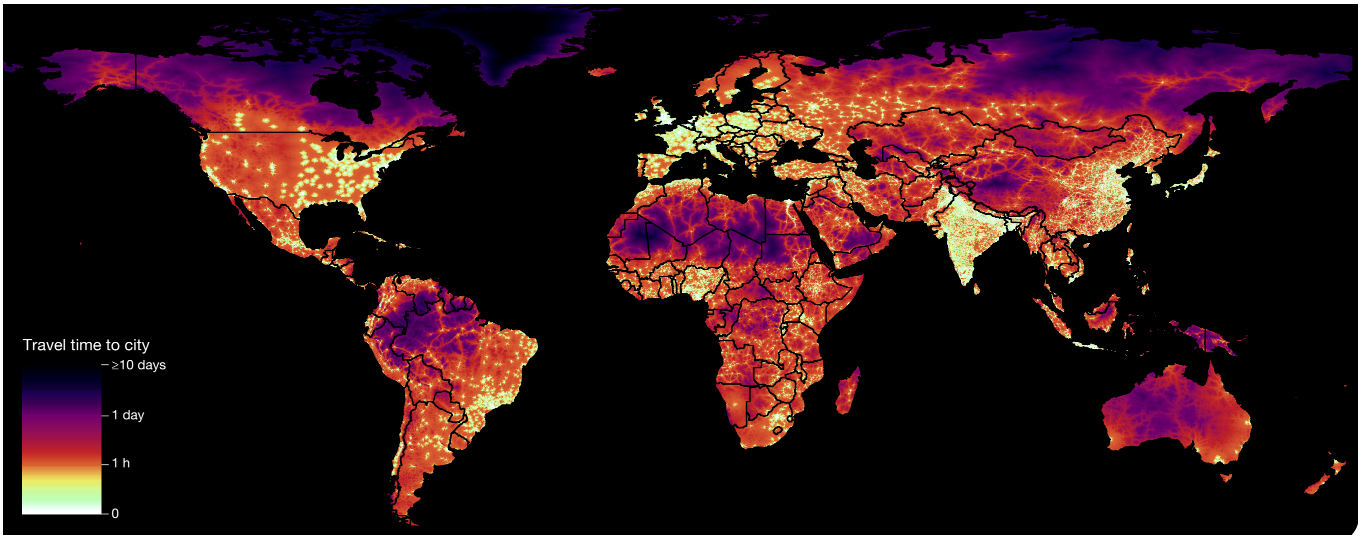


Figure 1 | Global map of travel time to cities for 2015. The accessibility map has a spatial resolution of approximately 1 × 1 km, spans 60° south to 85° north latitude, and enumerates travel time to the city with the shortest associated journey.

was the inclusion of minor roads (for example, unpaved rural roads and exurban residential streets). The OSM database also provided the necessary information for assigning country- and road-type-specific speed data to road pixels. The Google roads data provided information critical for maintaining connectivity in areas where OSM coverage was sparse and/or fragmented owing to its piecemeal data collection approach. All input datasets were combined to create a global 'friction surface' at a resolution of approximately 1 by 1 km at the equator (that is, 30-arcsec resolution), effectively enumerating the generalized rates at which humans can move through each pixel of the world's surface.

The Global Human Settlement Grid of high-density land cover (GHS-HDC)[7] was used to represent cities for this research. This dataset consisted of 13,840 urban areas, an increase of 62.6% from the city points dataset used for the 2000 map[2]. We applied an algorithm that identified, for each pixel, the path with the least cost[12] through the friction surface to any pixel defined as an urban area within the GHS-HDC. This approach ultimately produced a global accessibility map enumerating travel time to the closest city for all areas between 60° south and 85° north latitude (Fig. 1). We generated the friction surface and most of the accessibility map using the Google Earth Engine platform[13], and by freely distributing our data and code our design supports the construction of bespoke accessibility maps for specific policy or programmatic priorities (for example, travel time to healthcare facilities, schools, employment centres, or markets). As with any global mapping effort, however, this study is subject to limitations that we describe in the Methods.

Our accessibility map illustrates broad patterns of accessibility globally, capturing both the asymmetric distribution of cities and vast inequalities in infrastructural development. Highly accessible areas include those with abundant transport infrastructure and/or many spatially disaggregated cities, suggesting that accessibility to cities can be increased by improvements in infrastructure as well as polycentric urban development. Further exploration of accessibility relative to gridded population datasets[14–17] shows that 80.7% of people (5.88 billion individuals) reside within one hour of cities, but accessibility is not equally distributed across the development spectrum. This disparity is evident when comparing accessibility for populations subdivided by World Bank classifications for income group and geographical region[18] (Fig. 2), as 90.7% of people in high-income countries (concentrated in Europe and North America) live within one hour of a city compared to 50.9% of people in low-income countries (concentrated in sub-Saharan Africa). The relationship between national wealth and accessibility is more ambiguous for upper- and lower-middle income countries owing to the high population and large number of spatially diffuse urban centres found in northern India. This finding illustrates differences in accessibility that can readily be discerned from our map

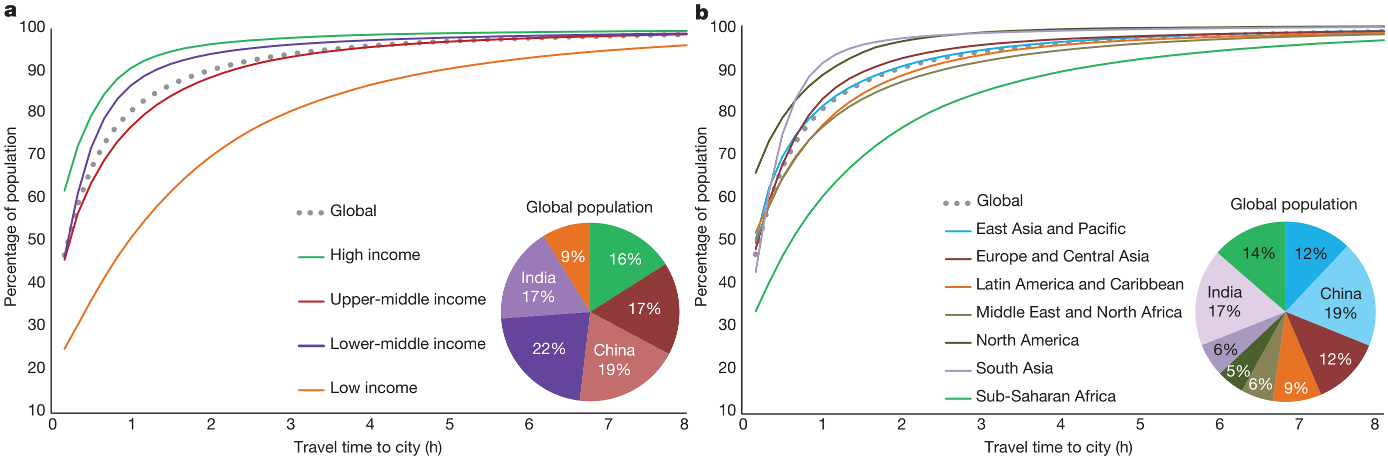


Figure 2 | Global disparities in accessibility. a, **b**, Travel time of global populations grouped by World Bank income categories (**a**) and regions (**b**). Lines show populations aggregated in 10-min increments and then divided by the total population of each group. The inset charts show the percentage of the global population within each group.

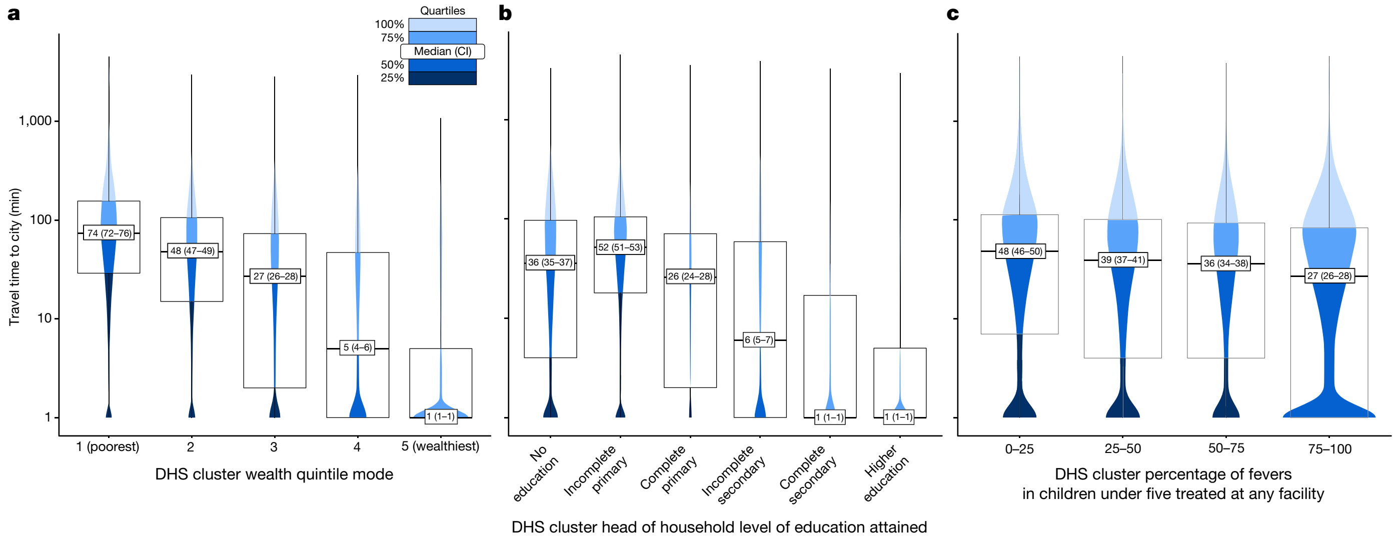


Figure 3 | Relating accessibility to human wellbeing. log-scale violin plots showing travel times (plus one minute) for household clusters in relation to metrics of wealth (**a**, $n = 47{,}761$), educational attainment (**b**, $n = 59{,}686$), and healthcare utilization (**c**, $n = 39{,}014$). The overlaid box plot hinges and colour-coding indicate the data quartiles, whiskers extend to the range of the data, violin shapes depict the data distributions, and medians and confidence intervals (CI = median ± 1.58(interquartile range/$n^{0.5}$)) are displayed at the box plot centres.

because it was produced using a globally consistent (and thus comparable) methodology. Although global summaries are informative, the accessibility map also supports fine-grained summarization and analysis (Extended Data Figs 1–3, Supplementary Information). Our map also provides a means for more nuanced characterizations of accessibility within rural populations than those that have been based upon commonly-used datasets such as binary classifications of urban versus rural land cover[19] or national estimates of urban population percentages[20] (Extended Data Fig. 4). As such, our map can be used to highlight major development gaps between predominantly urban and rural populations and provide a means of enumerating accessibility within rural populations along a continuum.

To analyse subnational relationships between accessibility and socioeconomic, educational, and health measures, we used data collected by the Demographic and Health Surveys (DHS) program between 2000 and 2015. Our DHS database consisted of 66,768 household clusters, from 122 unique surveys spanning 52 countries, which were aggregated from nearly 1.77 million surveyed households. Although DHS surveys primarily cover low-to-middle-income countries, and thus do not fully represent all socioeconomic contexts, our results illustrate reciprocal relationships between accessibility and key metrics of human wellbeing in many geographical, political, and economic settings. We found a clear association between higher household wealth and greater accessibility to population centres (Fig. 3a). Similar patterns emerged for measures of educational attainment (Fig. 3b) and treatment of fever among children under five (Fig. 3c). Although exceptions occurred (for example, there were wealthy household clusters far from cities and vice versa), and the selected metrics were strongly collinear, the association between accessibility to cities and indicators of human wellbeing in low-to-middle-income countries was unequivocal.

The 2015 global accessibility map provides a high level of detail while also characterizing spatial heterogeneity in accessibility at a range of spatial scales. Our map is likely to serve as a critical input for future geospatial modelling endeavours, including those that highlight positive aspects of low accessibility, such as the protective effect that remoteness provides to wilderness areas[21,22], or reinforce the need for strategic road building that avoids unnecessary environmental damage[23,24]. We illustrate such use through a cursory case study relating accessibility with forest loss between 2000 and 2015 (Extended Data Figs 1–3) based upon the Global Forest Change dataset[25]. This study shows the potential of our map for contributing to natural science research, conservation efforts, and formulation of environmental policy.

While results from our exploratory analyses do not causally link accessibility to metrics of development (for example, they cannot be used to determine whether places are more affluent because of greater accessibility or vice versa), they nevertheless illustrate the relationship between travel time and socioeconomic outcomes encompassed within the Sustainable Development Goals. Many view reducing inequalities in the accessibility of the services, institutions, and economic opportunities offered by cities as a vital pathway to sustainable development and improved livelihoods for all populations. Our analysis supports this perspective and future studies should track and evaluate the multifaceted effects that result from improved accessibility.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 May; accepted 23 November 2017.
Published online 10 January 2018.

1. United Nations. *Transforming our World: The 2030 Agenda for Sustainable Development.* (United Nations Department of Economic and Social Affairs, 2015).
2. Nelson, A. Travel time to major cities: a global map of accessibility. http://forobs.jrc.ec.europa.eu/products/gam/ (Global Environment Monitoring Unit, Joint Research Centre of the European Commission, 2008).
3. Young, A. Inequality, the urban–rural gap and migration. *Q. J. Econ.* **128,** 1727–1785 (2013).
4. Fotso, J.-C. Urban–rural differentials in child malnutrition: trends and socioeconomic correlates in sub-Saharan Africa. *Health Place* **13,** 205–223 (2007).
5. Bloom, D. E., Canning, D. & Fink, G. Urbanization and the wealth of nations. *Science* **319,** 772–775 (2008).
6. Frelat, R. *et al.* Drivers of household food availability in sub-Saharan Africa based on big data from small farms. *Proc. Natl Acad. Sci. USA* **113,** 458–463 (2016).
7. Pesaresi, M. & Freire, S. GHS settlement grid following the REGIO model 2014 in application to GHSL landsat and CIESIN GPW v4-multitemporal (1975–1990–2000–2015) Data Sets. http://data.europa.eu/89h/jrc-ghsl-ghs_smod_pop_globe_r2016a (Joint Research Centre of the European Commission, 2016).
8. Nelson, A. & Chomitz, K. M. Effectiveness of strict vs. multiple use protected areas in reducing tropical forest fires: a global analysis using matching methods. *PLoS ONE* **6,** e22722 (2011).
9. Schmitz, C. *et al.* Trading more food: implications for land use, greenhouse gas emissions, and the food system. *Glob. Environ. Change* **22,** 189–209 (2012).
10. Gilbert, M. *et al.* Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across Asia. *Nat. Commun.* **5,** 4116 (2014).
11. Bhatt, S. *et al.* The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature* **526,** 207–211 (2015).
12. Dijkstra, E. W. A note on two problems in connexion with graphs. *Numer. Math.* **1,** 269–271 (1959).

13. Gorelick, N. *et al.* Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202,** 18–27 (2017).
14. Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P. & Tatem, A. J. High resolution population distribution maps for Southeast Asia in 2010 and 2015. *PLoS ONE* **8,** e55882 (2013).
15. Linard, C., Gilbert, M., Snow, R. W., Noor, A. M. & Tatem, A. J. Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS ONE* **7,** e31743 (2012).
16. Sorichetta, A. *et al.* High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Sci. Data* **2,** 150045 (2015).
17. Center for International Earth Science Information Network and Centro Internacional de Agricultura Tropical. Gridded Population of the World, Version 3 (GPWv3): Population Density Grids. http://dx.doi.org/10.7927/H4ST7MRB (NASA Socioeconomic Data and Applications Center, 2005).
18. World Bank. GDP (current US$). http://data.worldbank.org/indicator/NY.GDP.MKTP.CD (2016).
19. Center for International Earth Science Information Network, Columbia University Institute for Demographic Research, International Food Policy Research Institute, The World Bank & Centro Internacional de Agricultura Tropical. Global Rural–Urban Mapping Project, Version 1 (GRUMPv1): Settlement Points Revision 01. https://doi.org/10.7927/H4BC3WG1 (NASA Socioeconomic Data and Applications Center, 2016).
20. United Nations. *World Urbanization Prospects: The 2014 Revision, Highlights* (United Nations Department of Economic and Social Affairs, 2014).
21. Allan, J. R. *et al.* Recent increases in human pressure and forest loss threaten many Natural World Heritage Sites. *Biol. Conserv.* **206,** 47–55 (2017).
22. Ibisch, P. L. *et al.* A global map of roadless areas and their conservation status. *Science* **354,** 1423–1427 (2016).
23. Laurance, W. F. *et al.* A global strategy for road building. *Nature* **513,** 229–232 (2014).
24. Laurance, W. F. & Arrea, I. B. Roads to riches or ruin? *Science* **358,** 442–444 (2017).
25. Hansen, M. C. *et al.* High-resolution global maps of 21st-century forest cover change. *Science* **342,** 850–853 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. C. Alonso, G. Pacheco, E. Brett, A. Arenzana, and B. A. Laken for the development of http://roadlessforest.eu/map.html and D. Pfeffer and K. Twohig for formatting of the figures and references. Funding was provided by a Google Earth Engine Research Award entitled "Developing and validating an online accessibility mapping tool powered by Google Earth Engine" and the "Roadless Forest" project "Making efficient use of EU climate finance: Using roads as an early performance indicator for REDD+ projects" (European Parliament/European Commission Directorate General for Climate Action).

Author Contributions D.J.W., A.N., S.P. and P.W.G. came up with the research concept and designed experiments. D.J.W. drafted the manuscript. D.J.W., A.N., H.S.G., W.T., A.L., B.M. and U.D. prepared and supplied data. D.J.W. conducted the analyses. M.H. and E.P. implemented algorithms. D.J.W. and A.L. coordinated the project. A.N., H.S.G., N.F., T.C.D.L., R.E.H., K.E.B. and S.Bh. supported the analyses and interpretations. J.R. and S.Be. produced visualizations. All authors discussed the results and revised the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.J.W. (daniel.weiss@bdi.ox.ac.uk).

Reviewer Information *Nature* thanks J. Birkmann and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

To model the time required for individuals to reach their most accessible city, we first quantified the speed at which humans move through the landscape. For this, we built on previous work that had integrated a number of infrastructural, political, and environmental datasets within a geographic information system (GIS)-based model[2]. The principle underlying this work was that all areas on Earth, represented as pixels within a 2D grid, had a cost (that is, time) associated with moving through them that we quantified as a movement speed within a cost or 'friction' surface. We then applied a least-cost-path algorithm[12] to the friction surface in relation to a set of high-density urban points. The algorithm calculated pixel-level travel times for the optimal path between each pixel and its nearest city (that is, with the shortest journey time). From this work we ultimately produced two products: (a) an accessibility map showing travel time to urban centres, as cities are proxies for access to many goods and services that affect human wellbeing; and (b) a friction surface that underpins the accessibility map and enables the creation of custom accessibility maps from other point datasets of interest.

Accessibility mapping methodology. The datasets that we used to construct the friction surface characterize the spatial locations and properties of roads, railroads, rivers, bodies of water, topographical conditions (elevation and slope angle), land cover, and national borders. The datasets were converted into aligning grids with a 30 arcsec resolution, with the pixel values representing speeds of movement. The layers were then combined following the approach defined within an earlier accessibility mapping project[2] such that the fastest mode of transport took precedence. The only exception to this logic was national borders, for which a crossing-time penalty was superimposed with priority over all other layers. The borders dataset was created from a UN global administrative units layer (GAUL) such that each border segment had a unique numerical identifier. This approach supports setting border-specific crossing times via a lookup table, however usable data do not presently exist for universally defining this parameter. As such, we used a static, one hour crossing-time penalty for all borders other than those within the Schengen and the UK–Ireland common security zones. Note that for readability the travel speeds for other input layers are provided in km h^{-1}, but the actual units within the friction surface raster are minutes required to travel one metre.

With the exception of the rivers input, each of the datasets we used in this project and the methods we used to pre-process the data have improved considerably relative to those of the 2000 accessibility map that was produced in 2008[2]. Two road datasets were combined for this research. The first road input layer consisted of vector data extracted from the OSM database, which was created by a user community dedicated to producing open-source, geocoded datasets of infrastructural resources. The OSM dataset was converted into a grid matching the geographic resolution and extent of the eventual friction surface. In cases for which vector features of more than one road type were present within a single pixel, the road type with the highest associated travel speed took precedence. This rasterization procedure resulted in an integer grid in which pixel values corresponded to a single road type that was subsequently linked to a speed via a lookup table, which we also derived from the OSM database. The lookup table contains the country-specific mean travel speeds associated with each available road type, as derived from attributes linked to individual roads by the OSM user community. We used this lookup table approach rather than direct assignment of road speeds because such speed-of-travel information was infrequently assigned to road vectors within the OSM database. This limitation also necessitated the creation of a global default lookup table, which we created using mean values for each road type from all countries. We applied values from the default table in cases where a country had no speed limit records for one or more road types found within it.

The second, equally important source of road data was the Google distance to roads surface. This Google dataset was also global in extent, although China and the Korean Peninsula were omitted owing to data distribution limitations. To combine the two road datasets the Google distance to roads raster was first restricted to include only pixels with values of 500 m or less, thereby approximating the 1 × 1 km rasterization of the OSM road vectors. Unlike the OSM data the resulting Google roads raster lacked road-type information. As such the OSM road-type designation took precedence if both layers contained road information for a single pixel. Where only Google road data were available, the pixels were given the default integer value corresponding to the generic 'road' class from OSM. When creating the friction surface, all pixels from the combined roads raster were assigned the road travel speeds from the OSM-based lookup tables. For the lookup procedure, we also used a grid of administrative units to determine each pixel's country association.

The railroad input layer was also created from the rasterized OSM surface. Unlike the OSM roads data, however, the railroads were not differentiated by type within OSM and thus consisted of a single class with a uniform movement speed. The railroad speed used in this project was 24.3 km h^{-1}, which was the mean value assigned to railroad vectors extracted from the OSM database.

Three datasets were used to account for travel time by water within the friction surface. River travel time was added via a global set of navigable rivers rasterized from the CIA World Data Bank II vector rivers dataset[26], which was the only hold-over input variable from the circa 2000 accessibility mapping endeavour[2]. Other options available for characterizing major rivers were explored, most notably the HydroSHEDS[27] river network and Vector Map Level 0 (VMAP0)[28] datasets, but we ultimately concluded that reusing the original data was warranted as neither of the alternatives was discernibly better given their associated limitations and lack of detail about which rivers were navigable. For inland water bodies, we used a newly created global surface-water occurrence dataset[29], which we first aggregated from its native 30-m resolution to create a layer that enumerated the fraction of each pixel's area that was covered by water at the resolution of the friction surface. In this procedure, all 30-m pixels within the resulting fractional surface-water dataset that were classified as water at least 80% of the time were considered permanent water, as 80% was the lowest occurrence value that we observed within ocean pixels when screening the data. The resulting fractional surface-water layer was then converted into a binary surface in which only pixels that were completely covered by permanent water were coded as a body of water amenable to be crossed by boat. The final dataset relating to water was a land–sea mask, which was used to identify ocean pixels. The movement speeds assigned to the water types within the friction surface were 10 km h^{-1} for rivers and lakes and 19 km h^{-1} for oceans. The value for rivers was based on inland travel speeds reported in the UK, Ireland, and Australia[30]. The ocean value was the average speed obtained from over 142 million observations of ocean-going passenger ships collected from the Automatic Identification System (AIS) and the Voluntary Observing Ship (VOS) program[30].

For all pixels not covered by any of the water, road or railroad datasets, we derived a baseline speed of movement overland (that is, on foot) using the MODIS MCD12Q1 land cover product[31] in which we assigned each land cover type a travel speed from a lookup table. The lookup table was created by summarizing results from an online survey designed to crowd-source estimates of how long it takes individuals to traverse each land cover type. The survey consisted of representative photos and global maps of each land cover type. Respondents were asked to estimate the amount of time it would take them to travel one kilometre (or one mile) on foot through each land cover type. The survey received 407 complete responses and, after standardizing the distance units, the median values for the fifteen land cover classes within the survey (in units of km h^{-1}) were as follows: evergreen needleleaf forest = 3.24, evergreen broadleaf forest = 1.62, deciduous needleleaf forest = 3.24, deciduous broadleaf forest = 4.00, mixed forest = 3.24, closed shrublands = 3.00, open shrublands = 4.20, woody savannas = 4.86, savannas = 4.86, grasslands = 4.86, permanent wetlands = 2.00, croplands = 2.50, cropland/natural vegetation = 3.24, snow and ice = 1.62, and barren or sparsely vegetated = 3.00. The two land cover classes we excluded from the survey were (a) urban and built-up, which was given a speed of 5 km h^{-1}, but this value is almost never needed owing to the higher speed (and thus precedence) of roads that dominate urban landscapes at 1 × 1 km resolution, and (b) open water, which was given a speed of 1 km h^{-1}. The speed for the open water pixels was assigned using the rationale that if these pixels were not considered inland water within the water bodies layer (and would thus have received the inland water speed associated with boat travel) they were probably more akin to permanent wetland pixels that had a very high subpixel fraction of water to circumnavigate on foot. As such, all pixels that were classified as open water in the land cover layer but not as permanent water in the water bodies layer were given a speed half as fast as the crowd-sourced median speed for permanent wetlands of 2 km h^{-1}.

The land-cover-dependent travel speeds were then adjusted to take into account the effect of topographical properties. Topographical data-sets used in this analysis were produced from the Global Multi-resolution Terrain Elevation Dataset 2010 (GMTED2010), a derivative of the Shuttle Radar Topography Mission data and produced by USGS[32]. The adjustment that we applied to elevation accounts for decreasing atmospheric density (and thus available oxygen) with altitude, which closely parallels the drop in maximal oxygen consumption (that is, VO_2 max, a measure of optimal heart and lung function) and thus decreased the predicted travel speed as a function of altitude[33]. On the basis of the standard atmosphere calculation, equation (1) shows the adjustment factor that we associated with elevation (in metres). We treated slope angle (in degrees) similarly, as steep terrain slows humans' ability to traverse it on foot. For the slope adjustment, we used Tobler's Hiking Function[34] as shown in equations (2) and (3), with Tobler's walking speed capped to a maximum of 5 km h^{-1} and then divided by five to convert it into a fraction of maximum travel speed. The elevation and slope adjustment factors were subsequently multiplied by the land-cover-dependent travel speeds, thus lowering the speed of travel on foot and increasing the time required to traverse each associated pixel within the friction surface.

$$\text{Elevation adjustment factor} = 1.016e^{-0.0001072 \times \text{elevation}} \qquad (1)$$

$$\text{Tobler's walking speed} = 6e^{-3.5|\tan(0.01745\times\text{slope angle})+0.05|} \quad (2)$$

$$\text{Slope adjustment factor} = \text{Tobler's walking speed}/5.0 \quad (3)$$

The final input for the accessibility map was the dataset of urban land cover, which was created using a layer from the Global Human Settlement (GHS) project[7]. This dataset was produced using a combination of satellite imagery and census data to map the spatial distribution of urban areas across the globe. To be consistent with data used in previous accessibility mapping research[2], we selected the 'high-density centres' variant of the GHS dataset, which is defined as "contiguous cells with a density of at least 1,500 inhabitants per km^2 or a density of built-up greater than 50% and a minimum of 50,000 inhabitants". The dataset contained a total of 13,840 unique urban areas and, unlike the circa 2000 accessibility map in which cities were represented as single geographical points, cities extracted from the GHS consisted of a cluster of pixels, thus effectively representing urban areas as polygons. The switch from single-point to multiple-pixel representations of cities was operationalized by extracting each urban pixel's centre coordinates and then applying the least-cost-path algorithm to only points on edges of urban areas. Over 400,000 high-density urban points were processed in this manner, not including points from the urban interiors, which were ignored to reduce redundant processing and later masked to have travel times of zero in the resulting accessibility map.

The friction surface was created entirely within Google Earth Engine, which was also used to create the majority of the accessibility surface. In contrast to the process used to create the friction surface, deriving the accessibility map was very computationally intensive and required a more complex processing chain. Within Earth Engine, accessibility surfaces were generated using the cumulativeCost function[35], a least-cost-path function that was an experimental tool implemented specifically for this project but is now freely available within Earth Engine. By harnessing the computational power of the Google cloud-computing system the cumulativeCost function shortened the production time of the global accessibility surface from several months (when relying on local computing resources alone) to approximately two weeks. Despite reducing the production time substantially, the cumulativeCost function was still an evolving tool that was not yet capable of producing the global accessibility map in a single run or reliably producing output for latitudes above 60° if the friction surface was in geographical coordinates (that is, units of degrees latitude and longitude). As such, we created the global accessibility map by mosaicking a set of 31 tiles, 24 of which encapsulated the most computationally demanding areas and were generated within Earth Engine, and seven of which we created outside Earth Engine. The limitations of the least-cost-path function within Earth Engine at high latitudes were due to the nature of processing raster data stored in geographical coordinates because distances at high latitudes span far more degrees of longitude (and thus more pixels) than comparable distances at low latitudes. In order to parallelize computations efficiently, the Earth Engine cumulativeCost function required specification of a maximum search distance from the source points (that is, high-density urban land cover pixel centres), which we set to 1,500 km for most of the globe but reduced to 1,000 km in areas from 50° to 55° latitude owing to the afore-mentioned processing limitations at high latitudes. For latitudes above 50° we calculated accessibility tiles using the gdistance package in R[36], thus ensuring an overlapping area of five-degrees latitude and providing data with which to compare the output maps from the differing sources (pixel values in these areas proved to be almost identical). We also calculated accessibility times locally for very remote islands at lower latitudes that were beyond the 1,500 km search distance threshold from their closest cities. Lastly, the cumulativeCost function in Earth Engine could not account for wrapping at ±180° longitude, so we created an alternative version of the friction surface centred at this longitude and reprocessed approximately one-fifth of the globe outside of Earth Engine to ensure that any pixels that had their closest cities on the opposite side of this 'edge' were ascribed accurate travel times. We then mosaicked all of the tiles together by selecting the minimum travel times for all pixels that fell within overlapping portions of multiple tiles. The result of this mosaicking operation is the global accessibility map shown in Fig. 1.

Model validation. We validated the accessibility map by comparing the travel times derived from least-cost-path calculations based on the friction surface with corresponding estimates derived from driving directions application within Google Maps (that is, comparison to travel time estimates derived using a network distance algorithm). The data source we used for validation consisted of settlement points from the Global Rural–Urban Mapping Project (GRUMP)[19]. Point pairs linking small settlements (that is, those with populations under 50,000 inhabitants) with their nearest city (that is, settlements with populations over 50,000 inhabitants) were processed using the friction surface approach to produce travel times akin to those within the accessibility map. After receiving special permission to automate the process of querying the Google driving directions application programming interface (API), we acquired validation travel times for each of the point pairs. A total of 53,091 validation point pairs were available after removing all coordinate pairs the API could not match. This approach limited the validation to places that fell along road networks, which precluded an assessment of the map's accuracy in the most remote areas on Earth. However, the applied validation approach does thoroughly validate the map with respect to human populations as (a) most of the world's people live in close proximity to a road of some variety, and (b) named points along road networks that we tested were indicative, from an accessibility perspective, of other points near roads at unnamed locations.

The validation results were encouraging, with an R^2 of 0.66 and a mean absolute error (that is, the average difference between the travel times regardless of sign) of 20.7 min. The distribution of the differences between the travel times also matched our expectations as 86.5% of the point pairs had lower travel time estimates from the friction surface approach compared to the values derived from the Google API. We attributed this unequal distribution to the presence of roads within the OSM dataset that were not present within the Google data and were thus unknown to the Google API when it calculated travel time via the Google road network. Another factor that helped to explain the preponderance of lower travel times to cities derived using the friction surface was that it incorporated other forms of travel (for example, by water). This factor was particularly important for explaining point pairs with very large travel time disparities. Additional reasons why the friction surface approach tended to produce lower travel times relative to the Google API include (a) the abstraction of vector roads within raster space, which effectively shortened some roads by reducing their sinuosity; (b) the speed limit look-up table which assigned speeds to roads that may be unrealistically high (for example, if road conditions are poor); and (c) the friction surface approach assumed constant and optimal travel speeds, unlike the Google API that incorporated temporally varying delays related to traffic density (for example, rush-hour delays).

The geographical distribution of the 13.5% of point pairs for which the travel times estimated by the Google API were shorter than the corresponding times from the friction surface approach was heavily concentrated in China. This is noteworthy, because the Google roads dataset that we used to create the friction surface lacked road data for China. As such, this finding suggests that there were roads in China that the Google API used to estimate travel times that were unaccounted for within the friction surface (that is, not present within OSM). Because both the under- and overestimates were partially attributable to incomplete road network data from either OSM or Google, using a combination of these road data sources to produce the accessibility map represents a major strength of this research.

Description of map limitations. Almost all research projects that generate modelled data at global scales rely on assumptions, generalizations, and the use of best-available (even if suboptimal) datasets. An important example of this for our work is that the time it takes an individual to move through the landscape is mediated by far more factors than just infrastructure or landscape properties. Wealth, in particular, is a likely determinant of whether someone travels on foot rather than taking a vehicle and thus substantially affects accessibility on the level of the individual. As such, users are cautioned from assuming our travel time estimates are universally applicable. It should also be noted, however, that because the accessibility mapping system was developed within Earth Engine, alternative variants of the accessibility map (for example, a walking only travel time map) can readily be created.

Another caveat relates to transport by rail and water, and specifically how the least-cost-path algorithm is able to freely transition from these networks to roads or vice versa when, in reality, switching modes of transport typically requires a station or port. In our friction surface this reality is not reflected and thus the least-cost-path algorithm will occasionally utilize water and rail pixels unrealistically. Railroads are also problematic, because there is insufficient data within OSM to differentiate railroads by type and thus all railroads are assigned a relatively slow speed. As such, high-speed train travel is effectively absent from our map, although that point is largely moot when mapping accessibility to the nearest city as high-speed trains typically link large cities together (that is, to utilize such a network an individual is likely already within a city of 50,000 or more people) and are therefore similar to air travel within this context.

Including slope angle as an input layer also presents challenges because the level of detail inherent to topographical datasets depends on the spatial resolution of the elevation data used to generate such metrics. For example, data at a 1 × 1 km resolution can only reflect the slope angle at that resolution, and are likely to miss large changes in topographical relief (and thus slope angle) at finer resolutions. A related caveat is the isotropic handling of slope angle such that it will always slow down movement regardless of whether the least-cost-path is oriented uphill or downhill. The net result of these caveats is that the friction and accessibility surfaces are less reliable for off-road areas, and particularly in mountainous regions. It should also be noted that erroneous data within the global topographical dataset resulted in unrealistically high travel time estimates for a small cluster of pixels (that is, less than 50) in western Colombia.

Another known limitation of the accessibility map is that it ignores geopolitical conflicts, such as those currently occurring in Syria, where degraded infrastructure and other impediments to movement will greatly affect travel times. The relative ease with which a new friction surface can be generated using our methodology, however, would allow us to create a new friction surface and accessibility map that takes degraded infrastructure into account and thus identify areas affected by the reduced access to resources. Likewise, national borders are particularly challenging to incorporate into the accessibility map, because many borders are contested and/or unrecognized by the UN (for example, the border between Northern Cyprus and Cyprus) and thus not accounted for within the friction surface. A related challenge is borders that are effectively impermeable barriers to travel for most people (for example, the border between North and South Korea). As previously stated, there are simply no reliable data that quantify how long it takes to cross most land borders, much less the contested ones, and thus we applied a universal value that reflects the fact that most borders slow movement, particularly at road crossings, while avoiding any baseless assumptions. These factors highlight the need for better global data on border permeability and crossing times, particularly in light of ongoing policy changes related to transnational migrant flows.

Seasonal changes also present a major challenge when characterizing accessibility, particularly when they pertain to areas periodically inundated by water or covered by deep snow in which movement may be precluded for parts of the year and/or people may change their mode of transport (and thus their movement speed). Likewise, rare events such as floods and earthquakes can sever transportation links such as roads and bridges, thus markedly changing spatial accessibility patterns. Because the accessibility map was produced largely in Earth Engine, such modifications to transportation networks can be addressed by rapidly remaking the friction surface to reflect the changed reality on the ground. There are several crowd-sourced examples demonstrating how quickly such information can be collected and made available for analysis (for example, https://www.hotosm.org). A more common issue of temporal variability in accessibility pertains to public forms of transportation, which typically operate on schedules that produce delays in travel time as individuals wait for buses, trains, or ferries. Similarly, traffic congestion will slow travel times both predictably (for example, at rush hour or owing to construction) and unpredictably (for example, because of traffic accidents). As such, our accessibility should not be viewed as applicable at every moment, but rather a general estimate of accessibility during nominal travelling hours and in the absence of adverse conditions.

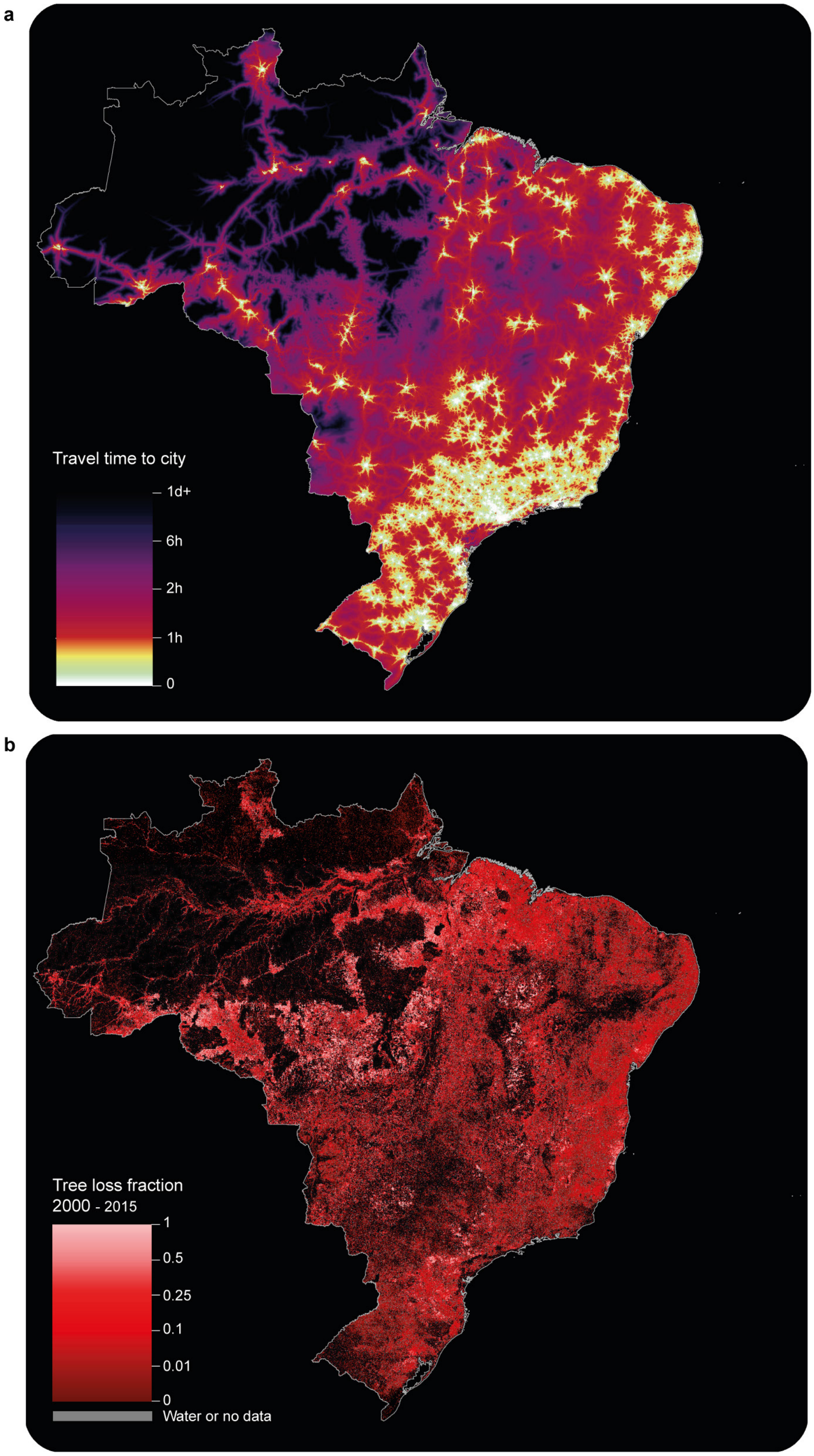
Exploratory analysis methodology: wealth, education, and healthcare utilization. The variables selected from the Demographic and Health Survey (DHS) Program database for exploratory analysis were household cluster measurements of the mode wealth index for heads of household, the mode educational attainment for heads of household, and the percentage of children receiving treatment for a fever (that is, healthcare utilization). The wealth and education variables were aggregated directly from questions asked within the surveys and owing to the categorical nature of these metrics, we selected the mode head of household values for analysis. The healthcare utilization metric, by contrast, was aggregated from individual-level data to provide cluster-level counts for both the numerator and denominator of the fever-treated fractions. For fever treatment this constitutes, respectively, the number of children (under five years of age) in each household cluster who received treatment for their fever divided by the total number of children within that household cluster who had a fever in the past two weeks. No statistical methods were used to predetermine sample size. Summaries of the DHS metrics relative to accessibility were depicted using violin plots (Fig. 3), which show the distribution and number of household clusters as the violin shape and area, respectively. To show the full data range these metrics were plotted using a logarithmic scale, which necessitated adding one minute to each survey cluster to plot those with travel times of zero. The added minute is reflected in Fig. 3, including the reported median and confidence interval values, the latter of which (derived as confidence intervals = median ± 1.58(interquartile range/$n^{0.5}$) (ref. 37)) are quite narrow as a result of the large sample sizes.

Exploratory analysis methodology: forest loss. Despite potential beneficial aspects of short travel times to cities for humans, higher accessibility has an associated environmental cost owing to the relative ease with which humans can extract natural resources in places closer to population centres. This relationship is observable when comparing a global dataset of changes in forest density from 2000 to 2015[25] with travel time to cities. The initial step in this analysis was to aggregate the data for forest coverage in 2000 and forest loss from 2000 to 2015 from their native 30-m resolutions to match the 30-arcsec resolution of the accessibility surface. This process resulted in two grids quantifying the fraction of each pixel that contained forest in 2000 and experienced forest loss by 2015. By multiplying these layers by a grid of area per pixel (to convert the results to km^2), binning (in 10-min increments) the pixels in both resulting layers according to their intersection with the accessibility map, and then dividing the binned totals for area of forest that experienced loss by the binned totals for area of forest in 2000, we obtained the summarized proportion of the original forest that experienced any level of loss for each travel time interval. The summarized results were then subdivided by country. Brazil and Indonesia were selected to illustrate the relationship between accessibility and forest loss using a combination of visual juxtaposition (Extended Data Figs 1, 2) and a comparison of summarizations of the national population, land area, and forest loss relative to accessibility (Extended Data Fig. 3). Our results indicate that forest loss tends to peak between 1 and 5 h from cities, or just beyond the relatively stable forest matrix associated with urban and suburban landscapes. This suggests that very close proximity to urban areas has a protective effect for forests, but a more nuanced assessment is that such areas were likely to have been heavily exploited before the year 2000 (that is, they had little readily harvestable forest remaining in 2000) and thus do not show major forest losses since the turn of the century. Geographical differences in both the magnitude and shape of the curves, however, reflect the importance of local context when interpreting results. Subnational patterns of forest loss also follow predictable patterns in both Brazil and Indonesia, with the least accessible forests showing the least amount of density loss.

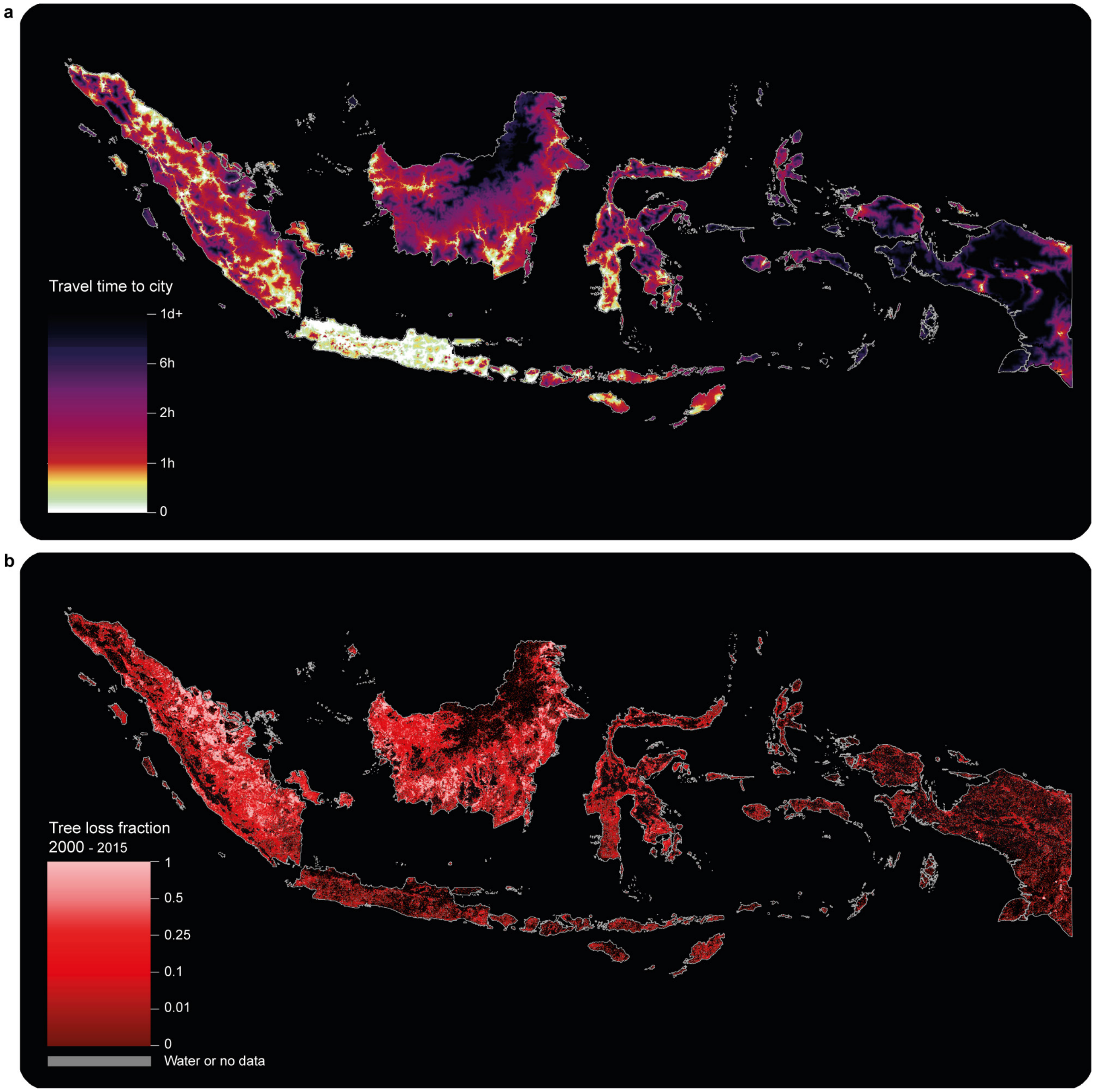
Code availability. The core code used to create the accessibility map is available for download from the Malaria Atlas Project website (https://www.map.ox.ac.uk/accessibility_to_cities/).

Data availability. The accessibility map is available for visualization and/or download at http://roadlessforest.eu/map.html and https://www.map.ox.ac.uk/accessibility_to_cities/.

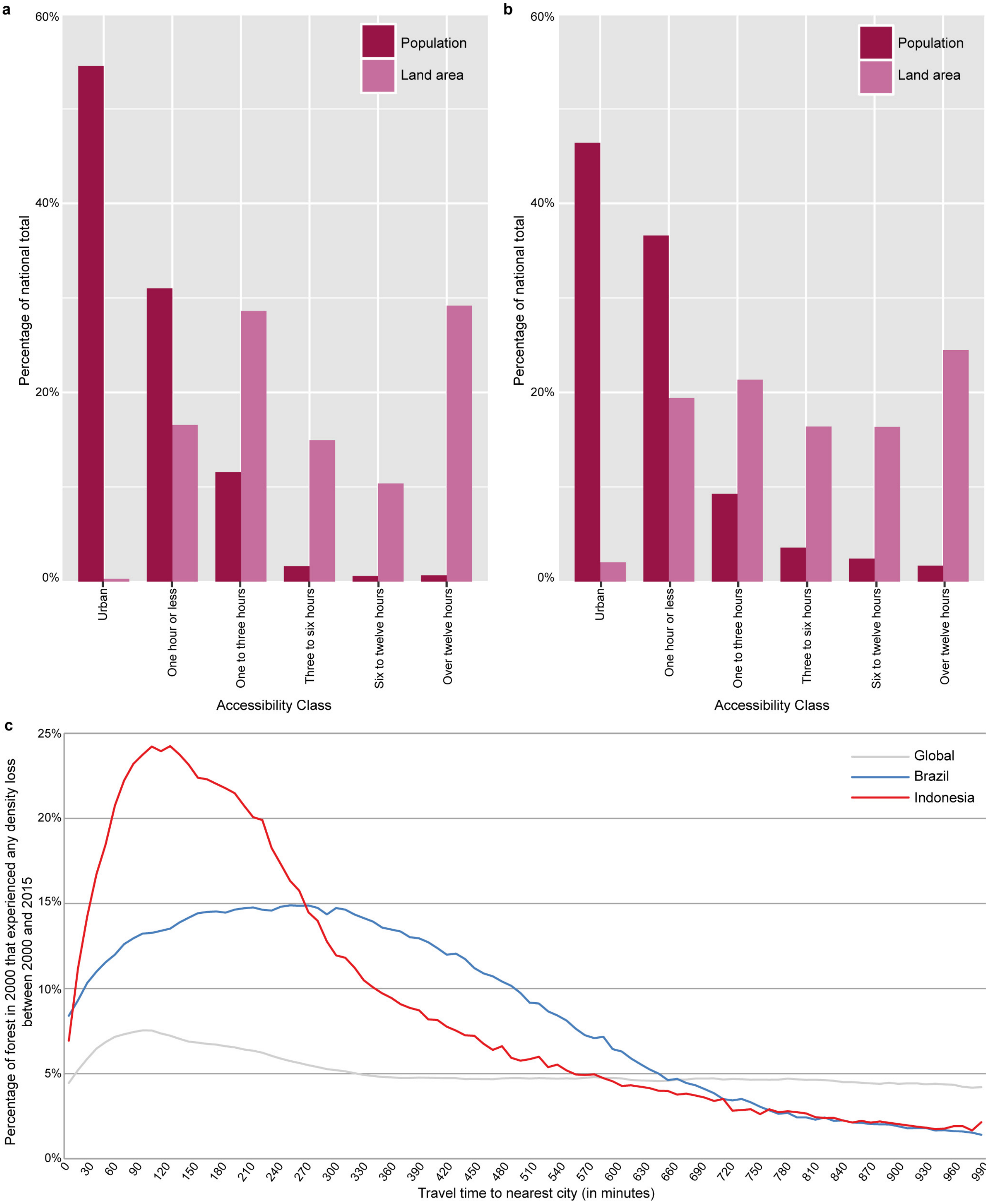
26. Central Intelligence Agency, Office of Geographic and Cartographic Research. World Data Bank II: North America, South America, Europe, Africa, Asia. https://doi.org/10.3886/ICPSR08376.v1 (Inter-university Consortium for Political and Social Research, 2000).
27. Lehner, B., Verdin, K. & Jarvis, A. New global hydrography derived from spaceborne elevation data. *Eos (Washington DC)* **89,** 93–94 (2008).
28. National Imagery and Mapping Agency. Vector Map Level 0 (VMAP0). http://www.mapability.com/info/vmap0_download.html (mapAbility, 1997).
29. Pekel, J.-F., Cottam, A., Gorelick, N. & Belward, A. S. High-resolution mapping of global surface water and its long-term changes. *Nature* **540,** 418–422 (2016).
30. Walbridge, S. *Assessing Ship Movements using Volunteered Geographic Information*, MA Thesis, Univ. California, Santa Barbara, (2013).
31. Friedl, M. A. *et al.* MODIS Collection 5 global land cover: algorithm refinements and characterization of new datasets. *Remote Sens. Environ.* **114,** 168–182 (2010).
32. Danielson, J. J. & Gesch, D. B. Global multi-resolution terrain elevation data 2010 (GMTED2010). https://explorer.earthengine.google.com/#detail/USGS%2FGMTED2010 (US Geological Survey, 2011).
33. Wehrlin, J. P. & Hallén, J. Linear decrease in. VO_{2max} and performance with increasing altitude in endurance athletes. Eur. *J. Appl. Physiol.* **96,** 404–412 (2006).
34. Tobler, W. *Three Presentations on Geographical Analysis and Modeling: Non-Isotropic Geographic Modeling; Speculations on the Geometry of Geography; and Global Spatial Analysis.* Technical Report 93-1. (National Center for Geographic Information and Analysis, 1993).
35. Google Earth Engine Developers. Cumulative Cost Mapping. https://developers.google.com/earth-engine/image_cumulative_cost (2017).
36. van Etten, J. R package gdistance: distances and routes on geographical grids. *J. Stat. Softw.* **76,** 1–21 (2017).
37. Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. *Graphical Methods for Data Analysis* (Wadsworth International Group, 1983).



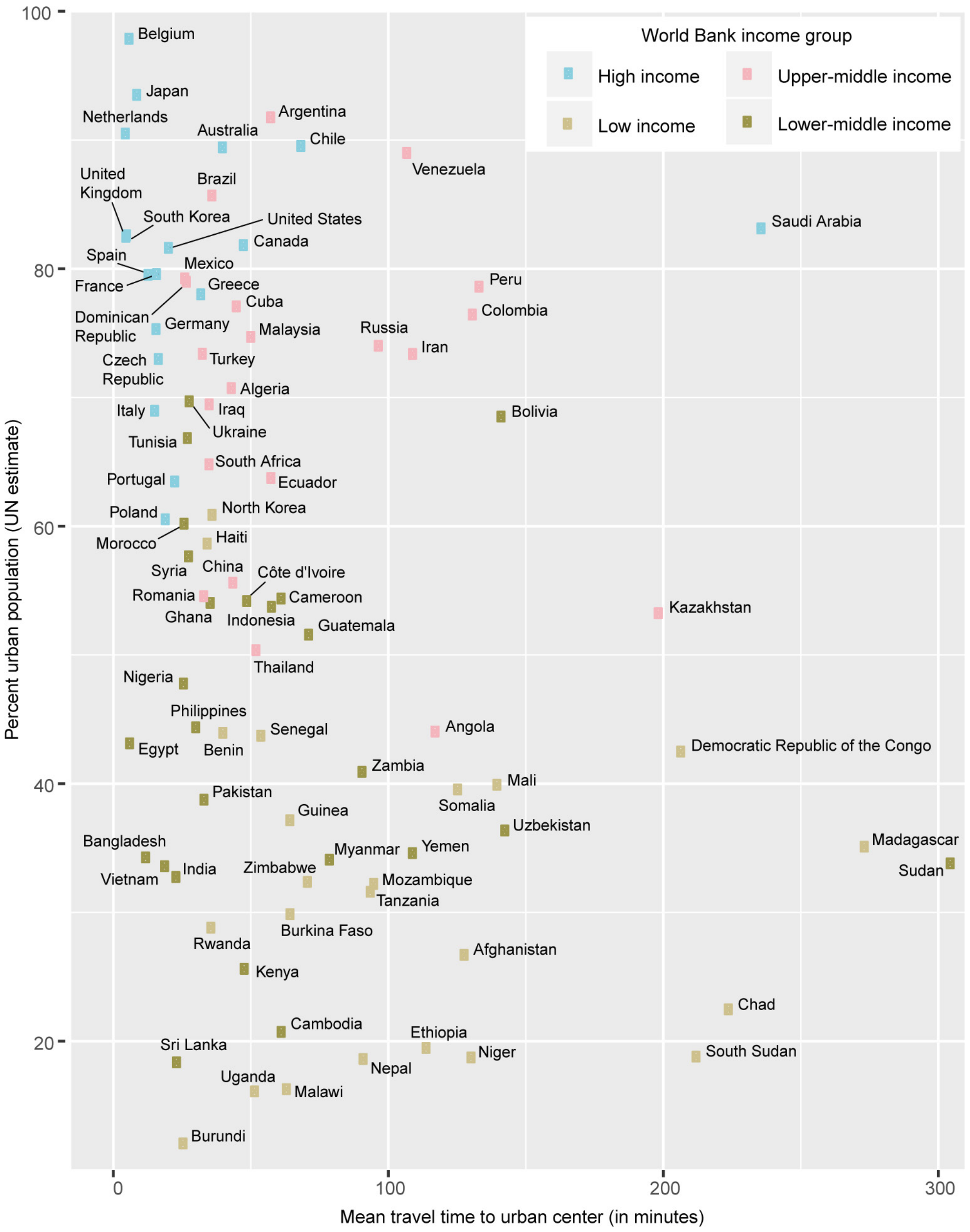
Extended Data Figure 1 | Accessibility and forest loss in Brazil. a, **b**, Maps of travel time to urban centres (**a**) and forest loss (**b**) from 2000 to 2015. Forest loss is defined as the fraction of land area identified as forest in 2000 that experienced any loss in forest density (but not necessarily total removal) by 2015.



Extended Data Figure 2 | Accessibility and forest loss in Indonesia. a, **b**, Maps of travel time to urban centres (**a**) and forest loss (**b**) from 2000 to 2015. Forest loss is defined as the fraction of land area identified as forest in 2000 that experienced any loss in forest density (but not necessarily total removal) by 2015.



Extended Data Figure 3 | Forest loss relative to accessibility. a, **b**, The distribution of the population and land area by accessibility category in Brazil (**a**) and Indonesia (**b**). **c**, The percentage of area that experienced any loss in forest density since 2000 for each country and the global average.



Extended Data Figure 4 | Travel time relative to percentage of urban population. Mean national accessibility for countries with populations over ten million relative to the percentage of urban population as estimated by the UN, colour-coded by World Bank income category.

doi:10.1038/nature25188

Paternal chromosome loss and metabolic crisis contribute to hybrid inviability in *Xenopus*

Romain Gibeaux[1], Rachael Acker[1], Maiko Kitaoka[1], Georgios Georgiou[2], Ila van Kruijsbergen[2], Breanna Ford[3], Edward M. Marcotte[4], Daniel K. Nomura[3], Taejoon Kwon[5], Gert Jan C. Veenstra[2] & Rebecca Heald[1]

Hybridization of eggs and sperm from closely related species can give rise to genetic diversity, or can lead to embryo inviability owing to incompatibility. Although central to evolution, the cellular and molecular mechanisms underlying post-zygotic barriers that drive reproductive isolation and speciation remain largely unknown[1,2]. Species of the African clawed frog *Xenopus* provide an ideal system to study hybridization and genome evolution. *Xenopus laevis* is an allotetraploid with 36 chromosomes that arose through interspecific hybridization of diploid progenitors, whereas *Xenopus tropicalis* is a diploid with 20 chromosomes that diverged from a common ancestor approximately 48 million years ago[3]. Differences in genome size between the two species are accompanied by organism size differences, and size scaling of the egg and subcellular structures such as nuclei and spindles formed in egg extracts[4]. Nevertheless, early development transcriptional programs, gene expression patterns, and protein sequences are generally conserved[5,6]. Whereas the hybrid produced when *X. laevis* eggs are fertilized by *X. tropicalis* sperm is viable, the reverse hybrid dies before gastrulation[7,8]. Here we apply cell biological tools and high-throughput methods to study the mechanisms underlying hybrid inviability. We reveal that two specific *X. laevis* chromosomes are incompatible with the *X. tropicalis* cytoplasm and are mis-segregated during mitosis, leading to unbalanced gene expression at the maternal to zygotic transition, followed by cell-autonomous catastrophic embryo death. These results reveal a cellular mechanism underlying hybrid incompatibility that is driven by genome evolution and contributes to the process by which biological populations become distinct species.

Hybrids produced through fertilization of *X. laevis* eggs with *X. tropicalis* sperm ($l_e \times t_s$) are viable, whereas *X. tropicalis* eggs fertilized by *X. laevis* sperm ($t_e \times l_s$) are not (Fig. 1a)[7,8]. Although cleavage divisions and rate of development of $t_e \times l_s$ hybrids were initially similar to *X. tropicalis* ($t_e \times t_s$) (Fig. 1b), hybrid embryos died abruptly as late blastulae and never initiated gastrulation. Before their death, hybrid embryos took on a deformed mushroom-like shape before lysing from the vegetal pole (Fig. 1c and Supplementary Video 1). Explants prepared from the opposite pole (animal caps) of mid-blastula $t_e \times l_s$ embryos also died within a few hours, indicating that embryo death is cell autonomous and not a result of faulty developmental cues (Fig. 1d and Supplementary Video 2). In contrast to $t_e \times l_s$ hybrids that die as embryos, haploid *Xenopus* embryos develop to the tadpole stage[8,9], suggesting that hybrid death is due to factors brought in by the *X. laevis* sperm to the *X. tropicalis* egg during fertilization. Irradiation of *X. laevis* sperm before fertilization, which destroys the DNA[10,11], resulted in a haploid phenotype (Fig. 1e and Supplementary Videos 3 and 4), indicating that $t_e \times l_s$ embryo death is due to the presence of the *X. laevis* genome. Cybrid embryos generated by irradiating *X. tropicalis* eggs, destroying the maternal DNA[8] before fertilization with *X. laevis* sperm, died before gastrulation similar to $t_e \times l_s$ embryos, indicating that hybrid inviability does not result from a conflict between the paternal and maternal genomes (Extended Data Table 1).

To visualize the dynamics of hybrid cell divisions, we injected mRNAs encoding fluorescent fusion proteins to label embryo chromosomes and mitotic spindles, and observed animal caps at early stage 9, which revealed anaphase defects and chromosome mis-segregation (Fig. 1f and Supplementary Video 5). Immunofluorescence of whole embryos confirmed the presence of lagging chromosomes and chromosome bridges in cells throughout hybrid blastulae, as well as the formation of micronuclei in interphase, whereas no such defects were observed in *X. tropicalis* embryos (Fig. 1g) or in the reverse viable hybrid (data not shown). Imaging of $t_e \times l_s$ embryos from stage 4 (eight cells) to stage 9 (thousands of cells) revealed micronuclei in 6–10% of the cells throughout hybrid development, but not in the *X. tropicalis* control (Extended Data Fig. 1a, b), indicating that chromosome mis-segregation in $t_e \times l_s$ hybrid embryos is unrelated to changes in gene expression at the onset of zygotic genome activation. Because the regular ploidy supported by the *X. tropicalis* egg is $N = 20$ chromosomes, but the $t_e \times l_s$ hybrid zygote must accommodate 28 chromosomes, we tested whether an increase in ploidy was causing chromosome mis-segregation and embryo death by applying a cold shock to *X. tropicalis* zygotes a few minutes after fertilization to suppress polar body extrusion and increase their ploidy to $N = 30$ chromosomes (Extended Data Fig. 1c). Micronuclei were not observed in cold-shocked embryos, which developed to the tailbud stage similarly to haploid embryos (Extended Data Fig. 1d). Thus, increasing the ploidy of *X. tropicalis* embryos does not cause chromosome mis-segregation or cell death, indicating a specific role for the *X. laevis* genome in hybrid inviability.

To determine whether assembly and function of the mitotic apparatus was affected, we used the *in vitro* egg extract system to examine spindle assembly and mitotic chromosome morphology. Metaphase-arrested *X. tropicalis* egg extract reconstituted spindle formation around nuclei isolated from stage 8 *X. tropicalis* ($N = 20$), *X. laevis* ($N = 36$), and viable hybrid embryos ($l_e \times t_s$; $N = 28$) (Fig. 2a). Spindle width scaled slightly with increasing genome size, but microtubule distribution was not affected by either genome size or content (Extended Data Fig. 1e), indicating that the presence of *X. laevis* DNA did not impair spindle assembly in *X. tropicalis* cytoplasm. To investigate chromosome morphology, *X. laevis* sperm nuclei were cycled through S phase in either *X. laevis* or *X. tropicalis* egg extract, induced to arrest in metaphase, and then stained with a DNA dye and antibodies to either CENP-A, the core centromeric histone variant, or Ndc80, an outer kinetochore component essential for linking centromeres to spindle microtubules[12]. Two fluorescent spots per chromosome were often visible in either extract, suggesting that the *X. tropicalis* extract is capable of replicating the *X. laevis* genome to generate duplicated

[1]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [2]Radboud University, Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, 6500 HB Nijmegen, The Netherlands. [3]Departments of Chemistry and Nutritional Sciences and Toxicology, University of California, Berkeley, California 94720, USA. [4]Department of Molecular Bioscience, Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, Texas 78712, USA. [5]Department of Biomedical Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea.

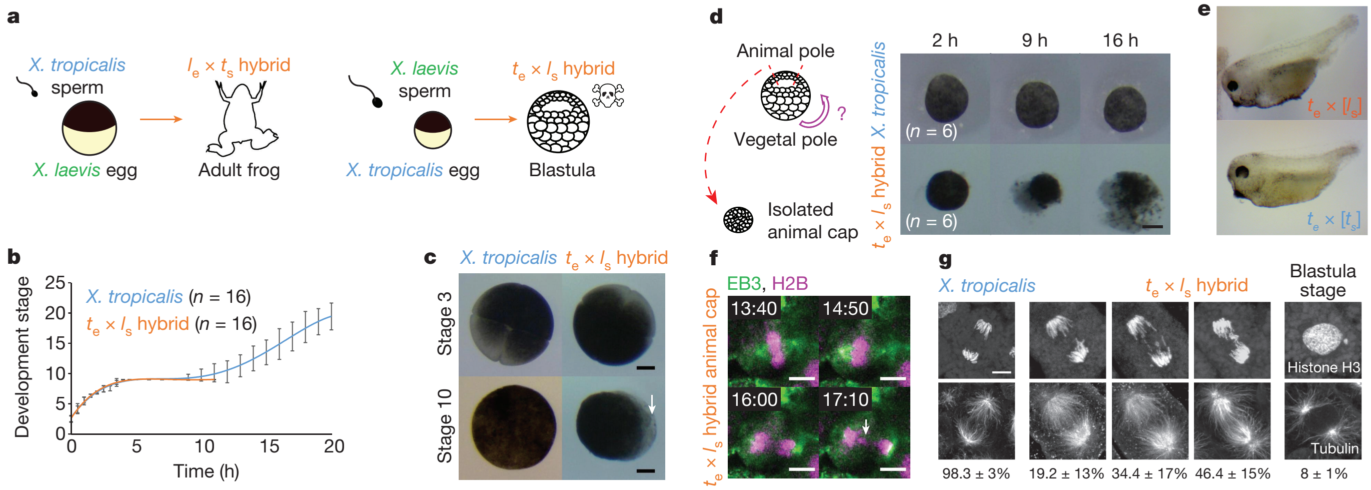


Figure 1 | Role of the *X. laevis* genome in $t_e \times l_s$ hybrid embryo death. **a**, Schematic of *X. laevis* and *X. tropicalis* cross-fertilization outcomes. **b**, Developmental timing in *X. tropicalis* and $t_e \times l_s$ hybrid embryos. Average is plotted for each time point. Error bars, s.d. **c**, Representative images of *X. tropicalis* and $t_e \times l_s$ hybrid embryos at stages 3 and 10 from experiments in **b** ($n = 16$ *X. tropicalis* and $n = 16$ $t_e \times l_s$ hybrid embryos from four independent experiments). Arrow indicates vegetal cells where death initiates. **d**, Schematic of animal cap assay and images of at 2, 9, and 16 h after isolation. Six animal caps were imaged and identical results were obtained in three different experiments. Scale bars in **c** and **d**, 200 μm. **e**, Images showing haploid phenotype following fertilization of *X. tropicalis* eggs with ultraviolet-irradiated sperm. Identical results were observed in $n = 3$ experiments. **f**, Time-lapse images of dividing cell in a $t_e \times l_s$ hybrid animal cap (Supplementary Video 5). Arrow indicates a mis-segregated chromosome. Mis-segregated chromosomes were observed in $n = 3$ live $t_e \times l_s$ hybrid animal caps in three experiments. Time is in minutes:seconds. **g**, Immunofluorescence images showing chromosome bridges, mis-segregated chromosomes, and micronuclei throughout $t_e \times l_s$ hybrid embryos. Scale bars in **f** and **g**, 10 μm. Quantification of $n = 81$ *X. tropicalis* and $n = 78$ $t_e \times l_s$ hybrid anaphases in $n = 17$ and 16 embryos, respectively, from four datasets obtained from three experiments presented as averages ± 1 s.d., show a significant difference by Fisher's 2 × 3 contingency test ($P = 0$). Quantification of micronuclei in $t_e \times l_s$ hybrid embryos is detailed in Extended Data Fig. 1b.

sister chromatids. However, we observed 13.5% fewer CENP-A-labelled and 12% fewer Ndc80-labelled chromosomes in *X. tropicalis* extract compared with *X. laevis* extract (Fig. 2b), suggesting that approximately two *X. laevis* chromosomes do not possess centromeres that become competent for kinetochore assembly following a cell cycle in *X. tropicalis* cytoplasm. Whole-genome sequencing of embryos at stage 9 before cell death revealed the specific loss of 228 megabases of *X. laevis* sequence from $t_e \times l_s$ hybrids (Fig. 2c), 96% of which was missing from just two chromosomes, 3L and 4L. By contrast, no genomic deletions were detected in viable $l_e \times t_s$ hybrid embryos

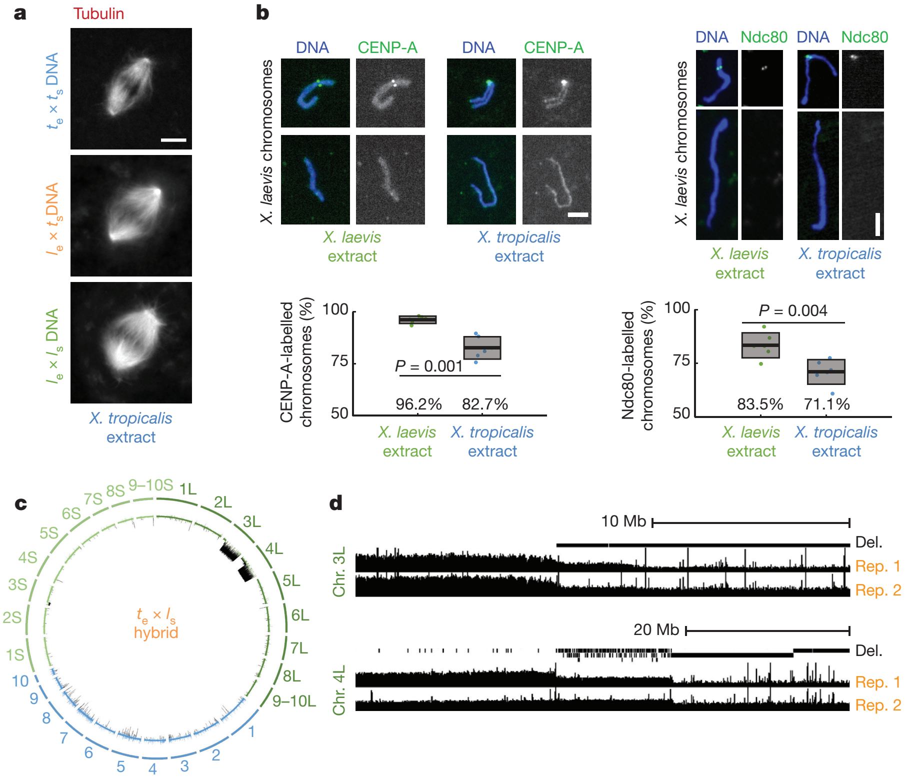


Figure 2 | Compatibility of *X. laevis* chromosomes with *X. tropicalis* cytoplasm. **a**, Fluorescence images of spindles formed around *X. tropicalis*, $l_e \times t_s$ hybrid, and *X. laevis* chromosomes in *X. tropicalis* egg extract. Scale bar, 10 μm. Quantification for $n = 147$, 103, and 156 spindles quantified for *X. tropicalis*, $l_e \times t_s$ hybrids, and *X. laevis* embryo nuclei, respectively, from three different egg extracts, is presented in Extended Data Fig. 1e. **b**, Fluorescence images of *X. laevis* chromosomes stained for CENP-A or Ndc80 following replication in *X. laevis* or *X. tropicalis* egg extract. CENP-A and Ndc80 labelling was quantified from six experiments (three biological replicates in two technical replicates), a total of $n = 1{,}792$ and $n = 1{,}959$ chromosomes, respectively, in *X. laevis* extract, and $n = 2{,}692$ and $n = 1{,}930$, respectively, in *X. tropicalis* extract. Scale bars, 5 μm. Box plots show the six experiment percentages as individual data points, their average as thick lines, and 1 s.d. as grey boxes. Ninety-five per cent confidence intervals are 96.2 ± 1.9% in *X. laevis* extract compared with 82.7 ± 5.7% in *X. tropicalis* extract for CENP-A, and 83.5 ± 6.1% compared with 71.1 ± 6.0% for Ndc80. *P* values were determined by two-tailed heteroscedastic *t*-test. **c**, Circle plot of whole-genome sequencing data for $t_e \times l_s$ hybrid embryos aligned and normalized to the genomes of *X. tropicalis* (blue) and *X. laevis* (green), with underrepresented genome regions in black. **d**, Expanded view of chromosome (Chr.) 3L and 4L breakpoints with deleted regions (Del.) indicated in two biological replicates (Rep.).

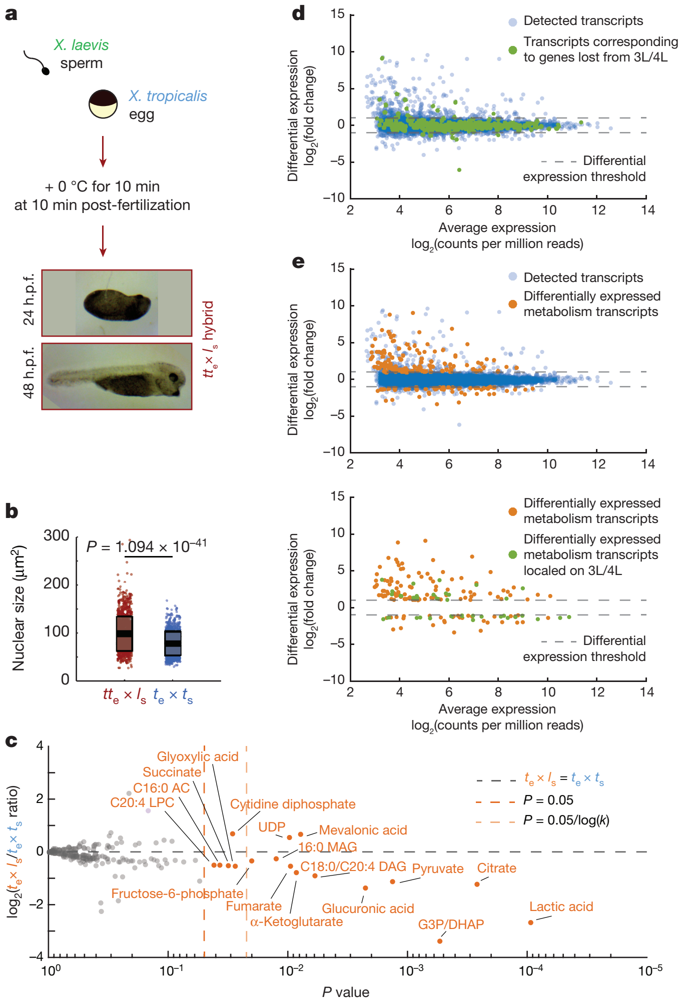


Figure 3 | Gene expression and metabolic changes preceding $t_e \times l_s$ hybrid embryo death. **a**, Schematic of polar body suppression experiment and images of $tt_e \times l_s$ rescued embryos 24 and 48 h.p.f. A total of nine $tt_e \times l_s$ embryos were obtained in four different experiments. **b**, Box plot of nuclear sizes ($n = 988$ nuclei from three $tt_e \times l_s$ embryos and $n = 777$ from three *X. tropicalis* embryos at stage 21) showing the average area as thick lines and 1 s.d. as grey boxes. Ninety-five per cent confidence intervals are $98.1 \pm 2.2\,\mu\text{m}^2$ for $tt_e \times l_s$ and $78.0 \pm 1.7\,\mu\text{m}^2$ for *X. tropicalis* embryos. *P* values were determined by two-tailed heteroscedastic *t*-test. **c**, Levels of 179 metabolites in *X. tropicalis* and $t_e \times l_s$ hybrid embryos 7 h.p.f. Levels were obtained from five samples from three independent fertilizations, each averaged and plotted as $\log_2$ of the ratio with the control (see Methods). *P* values were calculated using a two-tailed homoscedastic *t*-test. The average and 1 s.d. for the differentially represented metabolites are shown, and 95% confidence intervals given in Extended Data Fig. 3b. **d**, Differential gene expression between $t_e \times l_s$ and $t_e \times t_s$ (see Methods). All detected transcripts ($n = 8{,}379$) are plotted in blue. Transcripts corresponding to genes lost from chromosomes 3L and 4L ($n = 270$) are plotted in green. **e**, Differential expression of metabolism genes between $t_e \times l_s$ and $t_e \times t_s$ (see Methods). Differentially expressed metabolism transcripts ($n = 165$) are plotted in orange, all detected transcripts ($n = 8{,}379$) in blue (top), and differentially expressed metabolism transcripts located on chromosomes 3L and 4L ($n = 35$) in green (bottom).

(data not shown). Chromosome regions adjacent to breakpoints were heterogeneous in abundance (Fig. 2d), consistent with stochastic chromosome breakage and loss. Notably, major breakpoints localized to a gap in the genome assembly, indicating the presence of repetitive elements. Chromosome loss and partial deletion have been observed in non-viable hybrids in fish[13,14] and *Drosophila*[15], but the underlying mechanisms were unclear. Our results suggest that $t_e \times l_s$ hybrid incompatibility may be due to divergence of centromeric sequences, which are poorly characterized in *Xenopus* but known to evolve rapidly[16], or to other unidentified repetitive DNA elements that lead to chromosome instability and ultimately prevent kinetochore assembly on chromosomes 3L and 4L.

We next investigated the link between chromosome loss and $t_e \times l_s$ hybrid embryo death. Micronuclei in cancer cells accumulate DNA damage[17–19] and, in *Xenopus*, DNA damage was shown to trigger apoptosis at the onset of gastrulation[20]. As in cancer cells, micronuclei in $t_e \times l_s$ hybrid embryos often lost envelope integrity and contained damaged DNA (Extended Data Fig. 2a, b). However, $t_e \times l_s$ hybrid death did not resemble TdT-mediated dUTP nick end labelling (TUNEL)-positive apoptotic death induced by chemical inhibitors of DNA replication or protein synthesis in *X. tropicalis* embryos (Extended Data Fig. 2c, Extended Data Table 2 and Supplementary Videos 6 and 7). We hypothesized that chromosome loss could lead to cell death by affecting gene expression at zygotic genome activation. To assess the effects of blocking gene expression globally, we treated *X. tropicalis* embryos with the transcription initiation inhibitor triptolide and observed a phenotype reminiscent of the timing and manner of the catastrophic $t_e \times l_s$ hybrid embryo death, although lysis did not initiate from the vegetal side (Extended Data Table 2 and Supplementary Videos 8 and 9). To test whether altering gene dosage could rescue hybrid viability, we applied cold shock to the hybrid zygote to suppress polar body extrusion and introduce a second copy of the *X. tropicalis* genome. Although extremely inefficient, a total of nine triploid hybrid $tt_e \times l_s$ embryos were obtained in four separate experiments and survived to tailbud/tadpole stages (Fig. 3a). Rescued embryos possessed significantly higher DNA content than diploid *X. tropicalis* embryos at stage 21 (Fig. 3b), but whole-genome sequencing revealed that *X. laevis* DNA was eliminated by the tadpole stage (Extended Data Fig. 3a, Extended Data Table 3 and Supplementary Table 1). Our results link $t_e \times l_s$ hybrid inviability with altered gene expression that can be rescued with a second copy of the *X. tropicalis* genome, and indicate that $t_e \times l_s$ hybrid embryo inviability is caused by defects at the onset of zygotic genome activation, and not by DNA damage and apoptosis.

Because metabolite pools are known to become crucial before gastrulation[21], we subjected $t_e \times l_s$ hybrid embryos to metabolic profiling at 7 h post-fertilization (h.p.f.), just before the characteristic deformation preceding lysis. Levels of 17 out of 179 metabolites detected were significantly altered (Fig. 3c). Reduction in lactic acid, the final product of fermentation, and tricarboxylic acid cycle intermediates revealed that glycolytic metabolism was impaired in the cytoplasm of $t_e \times l_s$ hybrid embryos, which could in turn alter lipid metabolites including neutral lipids such as diacylglycerols and monoacylglycerols, as well as fatty-acid oxidation metabolites such as acyl carnitines (Extended Data Fig. 3b). While inhibition of mitochondrial ATP synthase led to cell cycle arrest at stage 9 (Supplementary Video 10), perturbing the early steps of glycolysis in $t_e \times t_s$ embryos induced cell death and lysis (Supplementary Video 11 and Extended Data Table 2). In particular, inhibition of glycogen phosphorylase to block the release of glucose from glycogen led to cell death at stage 9, initiating from the vegetal side of the embryo (Supplementary Video 12). These results are consistent with glycolytic defects as a primary cause of $t_e \times l_s$ hybrid embryo death. However, other defects that contribute to hybrid incompatibility could be masked by the abrupt cell lysis, such as conflicts between the paternal genome and maternal mitochondria[22,23].

To evaluate the link between the metabolic defects and specific chromosome loss, we used a statistical analysis[24] to classify the list of 1,803 genes mapped to the regions lost from chromosomes 3L and 4L in $t_e \times l_s$. We found that metabolic processes, particularly in glycolysis, were significantly over-represented (Extended Data Table 4). Transcriptome profiling of $t_e \times l_s$ hybrid embryos at 7 h.p.f. (Supplementary Table 2) revealed that although a large fraction of genes lost from chromosomes 3L and 4L were not differentially expressed compared with wild-type embryos (>92%; Fig. 3d), 27.1% of the differentially expressed genes related to metabolism (Fig. 3e, top), including

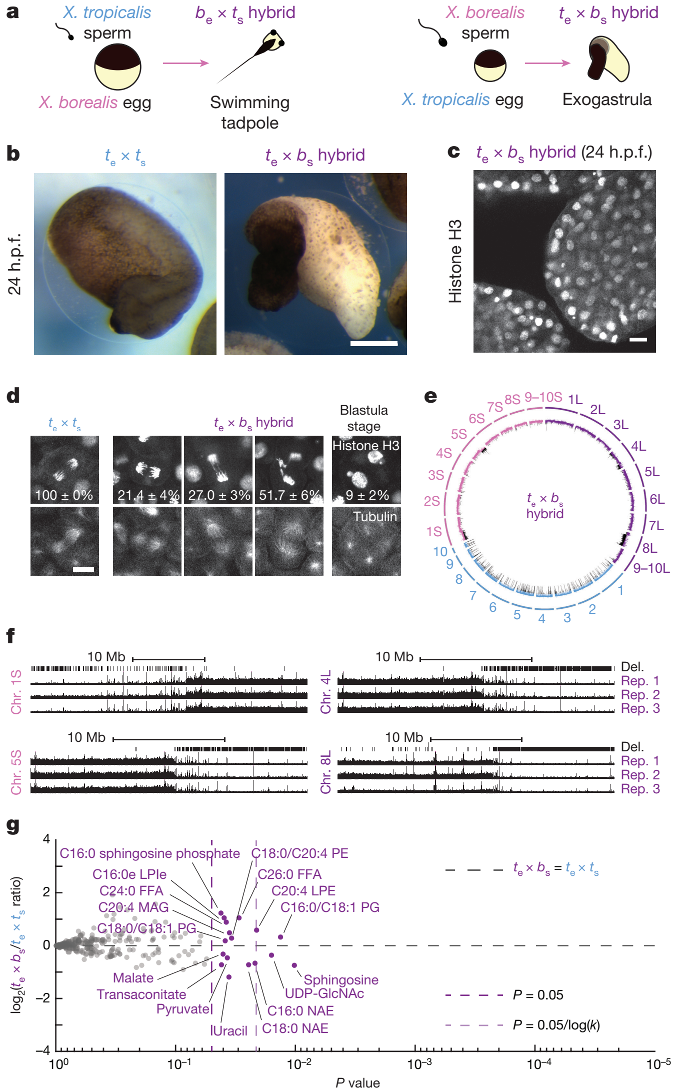


Figure 4 | Chromosomal loss in exogastrulating $t_e \times b_s$ hybrid embryos. a, Schematic of *X. borealis* and *X. tropicalis* cross-fertilization outcomes. **b**, Representative images of $t_e \times t_s$ compared with $t_e \times b_s$ embryos at 24 h.p.f. This result was reproduced in four separate experiments. Scale bar, 200 µm. **c**, Immunofluorescence image of $t_e \times b_s$ hybrid embryo at 24 h.p.f. showing nuclei and micronuclei. Similar defects at this stage were observed in six different embryos. **d**, Immunofluorescence images showing chromosome bridges, mis-segregating chromosomes, and micronuclei throughout $t_e \times l_s$ hybrid embryos. Scale bars, 20 µm. Quantification of $n = 33$ *X. tropicalis* and 63 $t_e \times b_s$ hybrid anaphases in $n = 6$ and 12 embryos, respectively, show a significant difference by Fisher's 2 × 3 contingency test ($P = 0$). Quantification of micronuclei in $t_e \times b_s$ hybrid embryos is detailed in Extended Data Fig. 4b. **e**, Circle plot of whole-genome sequencing data for $t_e \times b_s$ hybrid embryos aligned and normalized to the genomes of *X. tropicalis* (blue) and *X. borealis* (purple). Underrepresented genome regions (black) represent 9.674% of chromosome 4L, 74.66% of 8L, 4.71% of 1S, and 14.4% of 5S. **f**, Expanded view of chromosome 1S, 5L, 4L, and 8L breakpoints with deleted regions indicated in three biological replicates. **g**, Levels of 241 metabolites in *X. tropicalis* and $t_e \times b_s$ hybrid embryos 7 h.p.f. (see Methods). Levels were obtained from five samples from three independent fertilizations each, averaged and plotted as log$_2$ of the ratio with the control (see Methods). *P* values were calculated using a two-tailed homoscedastic *t*-test. The average and 1 s.d. for the differentially represented metabolites are shown, and 95% confidence intervals given in Extended Data Fig. 3c. Note that few metabolites are altered significantly and are distinct from those altered in $t_e \times l_s$ hybrids (see Extended Data Fig. 3b, c).

PDK1 (pyruvate dehydrogenase kinase). Moreover, 36.7% of the significantly under-expressed metabolism genes are found on chromosomes 3L and 4L (Fig. 3e, bottom), including *GFPT1* (fructose-6-phosphate aminotransferase) and *HPDL* (4-hydroxyphenylpyruvate dioxygenase).

To further characterize the specificity and mechanism underlying $t_e \times l_s$ hybrid incompatibility, we compared the outcome of cross-fertilizations between *X. tropicalis* and another allotetraploid *Xenopus* species, *X. borealis*[25]. Analogous to hybridization between *X. laevis* and *X. tropicalis*, we observed that *X. borealis* eggs fertilized with *X. tropicalis* sperm ($b_e \times t_s$) were viable, whereas the reverse hybrid ($t_e \times b_s$) was not (Fig. 4a). However, the $t_e \times b_s$ embryos did not lyse, but exogastrulated and survived for hours with intact cells (Fig. 4b, c and Supplementary Video 13). Similar to $t_e \times l_s$, $t_e \times b_s$ embryos displayed chromosome loss through anaphase defects and formation of micronuclei (Fig. 4d and Extended Data Fig. 4a–c). Notably, whole $t_e \times b_s$ hybrid genome sequencing revealed that, although the loss was specific for the paternal genome as in the $t_e \times l_s$ hybrid, specific regions of four different *X. borealis* chromosomes were affected (Fig. 4e, f, Extended Data Table 3 and Supplementary Table 1). Furthermore, metabolomics of $t_e \times b_s$ embryos revealed a distinct profile with less severe alterations than observed for $t_e \times l_s$ (Fig. 4g).

Altogether, our results indicate that hybrid instability in *Xenopus* results primarily from post-zygotic conflicts between the maternal cytoplasm and the paternal genome that lead to loss of specific genomic regions and downstream gene dosage defects. These findings highlight the role of genome evolution and transmission in defining hybrid fates and speciation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 November 2016; accepted 20 November 2017.
Published online 10 January 2018.

1. Seehausen, O. *et al.* Genomics and the origin of species. *Nat. Rev. Genet.* **15,** 176–192 (2014).
2. Presgraves, D. C. The molecular evolutionary basis of species formation. *Nat. Rev. Genet.* **11,** 175–180 (2010).
3. Session, A. M. *et al.* Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* **538,** 336–343 (2016).
4. Brown, K. S. *et al. Xenopus tropicalis* egg extracts provide insight into scaling of the mitotic spindle. *J. Cell Biol.* **176,** 765–770 (2007).
5. Hirsch, N., Zimmerman, L. B. & Grainger, R. M. *Xenopus*, the next generation: *X. tropicalis* genetics and genomics. *Dev. Dyn.* **225,** 422–433 (2002).
6. Yanai, I., Peshkin, L., Jorgensen, P. & Kirschner, M. W. Mapping gene expression in two *Xenopus* species: evolutionary constraints and developmental flexibility. *Dev. Cell* **20,** 483–496 (2011).
7. Bürki, E. The expression of creatine kinase isozymes in *Xenopus tropicalis*, *Xenopus laevis laevis*, and their viable hybrid. *Biochem. Genet.* **23,** 73–88 (1985).
8. Narbonne, P., Simpson, D. E. & Gurdon, J. B. Deficient induction response in a *Xenopus* nucleocytoplasmic hybrid. *PLoS Biol.* **9,** e1001197 (2011).
9. Hamilton, L. Androgenic haploids of a toad, *Xenopus laevis. Nature* **179,** 159 (1957).
10. Goda, T. *et al.* Genetic screens for mutations affecting development of *Xenopus* tropicalis. *PLoS Genet.* **2,** e91 (2006).
11. Wühr, M. *et al.* Evidence for an upper limit to mitotic spindle length. *Curr. Biol.* **18,** 1256–1261 (2008).
12. Cheeseman, I. M. The kinetochore. *Cold Spring Harb. Perspect. Biol.* **6,** a015826 (2014).
13. Fujiwara, A., Abe, S., Yamaha, E., Yamazaki, F. & Yoshida, M. C. Uniparental chromosome elimination in the early embryogenesis of the inviable salmonid hybrids between masu salmon female and rainbow trout male. *Chromosoma* **106,** 44–52 (1997).
14. Sakai, C. *et al.* Chromosome elimination in the interspecific hybrid medaka between *Oryzias latipes* and *O. hubbsi*. *Chromosome Res.* **15,** 697–709 (2007).
15. Ferree, P. M. & Barbash, D. A. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. *PLoS Biol.* **7,** e1000234 (2009).
16. Kalitsis, P. & Choo, K. H. A. The evolutionary life cycle of the resilient centromere. *Chromosoma* **121,** 327–340 (2012).
17. Crasta, K. *et al.* DNA breaks and chromosome pulverization from errors in mitosis. *Nature* **482,** 53–58 (2012).
18. Hatch, E. M., Fischer, A. H., Deerinck, T. J. & Hetzer, M. W. Catastrophic nuclear envelope collapse in cancer cell micronuclei. *Cell* **154,** 47–60 (2013).

19. Terradas, M., Martín, M., Tusell, L. & Genescà, A. DNA lesions sequestered in micronuclei induce a local defective-damage response. *DNA Repair (Amst.)* **8,** 1225–1234 (2009).
20. Hensey, C. & Gautier, J. A developmental timer that regulates apoptosis at the onset of gastrulation. *Mech. Dev.* **69,** 183–195 (1997).
21. Vastag, L. *et al.* Remodeling of the metabolome during early frog development. *PLoS ONE* **6,** e16881 (2011).
22. Ma, H. *et al.* Incompatibility between nuclear and mitochondrial genomes contributes to an interspecies reproductive barrier. *Cell Metab.* **24,** 283–294 (2016).
23. Lee, H. Y. *et al.* Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. *Cell* **135,** 1065–1073 (2008).
24. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.* **44** (D1), D336–D342 (2016).
25. Schmid, M. & Steinlein, C. Chromosome banding in Amphibia. XXXII. The genus *Xenopus* (Anura, Pipidae). *Cytogenet. Genome Res.* **145,** 201–217 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the Heald laboratory, present and past, for support and discussions. We thank the students who helped with some of the experiments: B. Castellano, J. Chen, S. Ramos, A. Sabillo, and K. Shih. We are also grateful to the Marine Biological Laboratory and the National *Xenopus* Resource for organizing the 2013 Advanced Imaging in *Xenopus* Workshop where several techniques used here were taught to R.G., and to J. Wallingford and A. Shindo for subsequent support. We thank the Welch, King, Harland, Rokhsar, Barton, and Fletcher laboratories at the University of California, Berkeley (UC Berkeley) for sharing reagents, materials, and expertise, as well as T. Stukenberg and A. Straight for providing us with the Ndc80 and CENP-A antibodies, respectively. We especially thank A. Mudd and D. Rokhsar for providing early access to the *X. borealis* genome assembly. This work used the Functional Genomics Laboratory, a QB3-Berkeley Core Research Facility at UC Berkeley as well as the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by National Institutes of Health (NIH) S10 OD018174 Instrumentation Grant. The confocal microscopy performed in this work was done at the UC Berkeley CRL Molecular Imaging Center, supported by National Science Foundation DBI-1041078. R.G. was initially supported by EMBO long-term fellowship ALTF 836-2013 and for most of this project by Human Frontier Science Program long-term fellowship LT 0004252014-L. R.A. was supported in part by a National Science Foundation REU Summer Fellowship in 2014. R.H. was supported by NIH R35 GM118183 and the Flora Lamson Hewlett Chair. D.K.N. was supported by NIH R01 CA172667. M.K. was supported by UC Berkeley Department of Molecular and Cell Biology NIH training grant 4T32GM007232-40. T.K. was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT and Future Planning (NRF-2016R1C1B2009302), and the UNIST Research Fund (grant number 1.160060.01). G.J.C.V., I.V.K., and G.G. were supported by NIH R01 HD069344.

Author Contributions R.H. and R.G. designed the project. R.G. performed the molecular, cell, and developmental biology experiments, aided by R.A., and analysed the data. M.K., together with R.G., performed the experiments related to *X. borealis* and analysed the data. G.J.C.V., I.V.K., and G.G. prepared and analysed the hybrid genomes. B.M. and D.K.N. performed the metabolomic profiling of hybrids. T.K. and E.M.M. contributed to the transcriptome data analysis. R.G. prepared the figures and wrote the manuscript with R.H., incorporating feedback from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to R.H. (bheald@berkeley.edu).

Reviewer Information *Nature* thanks E. Amaya and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Chemicals. Unless otherwise stated, all chemicals were purchased from Sigma-Aldrich.

Frogs. All animal experimentation in this study was performed according to the Animal Use Protocol approved by the UC Berkeley Animal Care and Use Committee. Mature *X. laevis*, *X. tropicalis*, and *X. borealis* frogs were obtained from NASCO, or the National *Xenopus* Resource (Woods Hole). Female *X. laevis* (1–4 years old), *X. tropicalis* (6 months to 4 years old), and *X. borealis* (2–3 years old) frogs were ovulated with no harm to the animals with 6-, 3-, and 4-month rest intervals, respectively. To obtain testes, males (same age ranges) were euthanized by over-anaesthesia through immersion in double-distilled (dd)H_2O containing 0.15% MS222 (tricaine) neutralized with 5 mM sodium bicarbonate before dissection, and then frozen at −20 °C.

Experimental design. No statistical methods were used to pre-determine sample size sufficient to generate statistically significant differences. All attempts at replication were successful. All experiments were performed independently at least three times (biological replicates). *Xenopus* frogs were selected randomly from our colony for ovulation and fertilization experiments. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

***In vitro* fertilization and cross-fertilization.** *X. laevis* males were injected with 500 U of human chorionic gonadotropin hormone (hCG) 12–24 h before dissection and testes were stored at 4 °C in 1× MR (100 mM NaCl, 1.8 mM KCl, 2 mM $CaCl_2$, 1 mM $MgCl_2$, and 5 mM HEPES–NaOH pH 7.6) for 1–2 weeks. *X. tropicalis* and *X. borealis* males were injected with 250 U and 300 U, respectively, of hCG 12–24 h before dissection, and testes were collected in Leibovitz L-15 Medium (Gibco, Thermo Fisher Scientific) supplemented with 10% fetal bovine serum (FBS; Gibco) for immediate use.

For *X. tropicalis* egg-based embryos, *X. tropicalis* females were primed with 25 U of hCG 12–24 h before use and boosted with 250 U of hCG on the day of the experiment. As soon as the first eggs were laid (~3 h after boosting), the *X. tropicalis* male was euthanized and dissected. Two *X. tropicalis* or *X. borealis* testes, or one-third of a *X. laevis* testis were each added to 1 ml of L-15 10% FBS. *X. tropicalis* females were squeezed gently to deposit eggs onto Petri dishes coated with 1.5% agarose in 1/10× MMR (1× MMR: 100 mM NaCl, 2 mM KCl, 2 mM $CaCl_2$, 1 mM $MgSO_4$, and 5 mM HEPES–NaOH pH 7.6, 0.1 mM EDTA). Testes were homogenized using scissors and a pestle in L-15 10% FBS. Any liquid in the Petri dishes was removed and the eggs were fertilized with 500 µl of sperm solution per dish. Eggs were swirled in the solution to separate them and incubated for 4 min with the dish slanted. Dishes were then flooded with ddH_2O, swirled, and incubated for 5–10 min. Buffer was exchanged for 1/10× MMR, the eggs incubated for 10 min, and jelly coats removed with a 3% cysteine solution (in ddH_2O–NaOH, pH 7.8). After extensive washing with 1/10× MMR (at least four times), embryos were incubated at 23 °C. At stage 2–3, fertilized embryos were sorted and placed in fresh 1/10× MMR within new Petri dishes coated with 1.5% agarose in 1/10× MMR.

For *X. laevis* egg-based embryos, *X. laevis* females were primed with 100 U of pregnant mare serum gonadotropin (PMSG, National Hormone and Peptide Program) at least 48 h before use and boosted with 500 U of hCG 14 h before the experiment. *X. laevis* females were squeezed gently to deposit eggs onto Petri dishes coated with 1.5% agarose in 1/10× MMR. Two *X. tropicalis* testes collected in L-15 10% FBS or one-third of a *X. laevis* testis were each added to 1 ml of ddH_2O and homogenized using scissors and a pestle. Any liquid in the Petri dishes was removed and the eggs were fertilized with 500 µl of sperm solution per dish. Eggs were swirled in the solution to individualize eggs as much as possible and incubated for 10 min. Dishes were flooded with 1/10× MMR, swirled, and incubated for 10–20 min. Jelly coats were removed with a 2% cysteine solution (in ddH_2O–NaOH, pH 7.8). After extensive washing (at least four times) with 1/10× MMR, embryos were incubated at 23 °C. At stage 2–3, fertilized embryos were sorted and placed in fresh 1/10× MMR in new Petri dishes coated with 1.5% agarose in 1/10× MMR.

For *X. borealis* egg-based embryos, *X. borealis* females were primed with 60 U of PMSG at least 48 h before use and boosted with 300 U of hCG 14 h before the experiment. Frogs were kept at 16 °C in 1/2× MMR. Eggs were picked from the tub and deposited onto Petri dishes coated with 1.5% agarose in 1/10× MMR. Two *X. tropicalis* or *X. borealis* testes were collected and homogenized using scissors and a pestle in L-15 10% FBS. Any liquid in the Petri dishes was removed and the eggs were fertilized with 500 µl of sperm solution per dish. Eggs were swirled in the solution to individualize eggs as much as possible and incubated for 10 min. Dishes were flooded with 1/10× MMR, swirled, and incubated for 10–20 min. Jelly coats were then removed with a 3% cysteine solution (in ddH_2O–NaOH, pH 7.8). After extensive washing (more than four times) with 1/10× MMR, embryos were incubated at 23 °C. At stage 2–3, fertilized embryos were sorted and placed in fresh 1/10× MMR in new Petri dishes coated with 1.5% agarose in 1/10× MMR.

All embryos were staged according to ref. 26.

Embryo chemical treatments and video imaging. Chemical treatments were performed in Petri dishes coated with exactly 5 ml of 1.5% agarose in 1/10× MMR covered with 10 ml 1/10× MMR for either regular incubations or video imaging for consistency. Concentrations were calculated relative to the covering volume of 1/10× MMR; no dilution within the volume in the agarose was assumed. Cycloheximide was added at a concentration of 0.1 $\mathrm{mg\,ml^{-1}}$ at stage 6.5 from 8 $\mathrm{mg\,ml^{-1}}$ stock in dimethylsulfoxide (DMSO). Hydroxyurea (Thermo Fisher Scientific) was added at a concentration of 30 mM at stage 3 from 600 mM stock in ddH_2O. Triptolide was added at a concentration of 25 µM at stage 2 from 25 mM stock in DMSO. Oligomycin was added at a concentration of 40 µM at stage 2 from 40 mM stock in DMSO. AP-III-a4 was added at a concentration of 30 µM at stage 2 from 1 mM stock in DMSO. Iodoacetic acid was added at a concentration of 50 mM at stage 2 from 1 M stock in ddH_2O. CP-91,149 was added at a concentration of 270 µM at stage 2 from 30 mM stock in DMSO. Corresponding volumes of DMSO or ddH_2O were added to controls.

Imaging dishes were prepared using an in-house PDMS mould designed to print a pattern of 0.9 mm large wells in agarose that allowed us to image six *X. tropicalis* embryos simultaneously within the 3 mm × 4 mm camera field of view for each condition. Embryos were imaged from stage 2–3. Treatment and control videos were taken simultaneously using two AmScope MD200 USB cameras (AmScope), each mounted on an AmScope SE305R stereoscope. Time-lapse movies were acquired at a frequency of one frame every 10 s for 20 h and saved as Motion JPEG using a MATLAB (The MathWorks) script. Movie post-processing (cropping, concatenation, resizing, and addition of scale bar) was done using MATLAB and Fiji[27]. All MATLAB scripts written for this study are available upon request. Two of the scripts used here were obtained through the MATLAB Central File Exchange: 'videoMultiCrop' and 'concatVideo2D' by 'Nikolay S'.

Embryo ploidy manipulations. To generate *X. tropicalis* haploid embryos ($t_e \times [t_s]$ and $t_e \times [l_s]$), fertilizations were conducted as detailed above with slight modifications to accommodate sperm ultraviolet-irradiation. Two *X. tropicalis* testes or one-third of a *X. laevis* testis were each added to 1.1 ml of L-15 10% FBS. Testes were homogenized using scissors and a pestle, and the solutions spun briefly using a benchtop centrifuge to pellet the tissue. One millilitre of supernatant was transferred into a glass Petri dish and irradiated within a Stratalinker UV-Crosslinker (Stratagene) with 50,000 µJ for *X. tropicalis* sperm or 2× 30,000 µJ for *X. laevis* sperm, swirling the solution in between the two irradiations. *X. tropicalis* eggs freshly squeezed onto Petri dishes coated with 1.5% agarose in 1/10× MMR were then fertilized with 500 µl of irradiated sperm solution per dish and processed as described above.

To generate $[t_e] \times l_s$ cybrid embryos and the haploid $[t_e] \times t_s$ controls, fertilizations were conducted as detailed above with slight modifications to accommodate for the ultraviolet-irradiation of the eggs. Two *X. tropicalis* testes or two-thirds of a *X. laevis* testis were each added to 1.1 ml of L-15 10% FBS. *X. tropicalis* females were squeezed gently to deposit eggs onto Petri dishes coated with 1.5% agarose in 1/10× MMR. Excess liquid was removed, eggs were swirled with a pestle to form a monolayer of properly oriented eggs, and immediately irradiated in a Stratalinker UV-Crosslinker (Stratagene) 2× with 40,000 µJ. Testes were homogenized using scissors and a pestle during the irradiation of the eggs. As soon as they were irradiated, the eggs were fertilized with 500 µl of sperm solution per dish and processed as described above.

To prevent polar body formation in either $tt_e \times t_s$ or $tt_e \times l_s$ experiments, fertilizations were conducted as detailed above with slight modifications to accommodate cold treatment. Fertilizations were performed within dishes coated with only 1–1.5 ml, instead of 5 ml, of 1.5% agarose in 1/10× MMR to accelerate cooling. Following the 4-min incubation with sperm, dishes were flooded with ddH_2O, swirled, and incubated for exactly 5 min. Buffer was then exchanged for ice-cold 1/10× MMR, the dishes transferred into a pipette tip box lid placed in a slushy ice bucket, and the eggs incubated for 10 min. The dishes were then removed from the bucket and the cold buffer was exchanged for RT 1/10× MMR. After 20 min, the jelly coat was removed with a 3% cysteine solution (in ddH_2O–NaOH, pH 7.8) and the embryos processed as described above.

Animal cap assay. At stage 8, embryos were placed in Danilchik's for Amy Medium (DFA medium; 53 mM NaCl, 5 mM Na_2CO_3, 4.5 mM potassium gluconate, 32 mM sodium gluconate, 1 mM $CaCl_2$, 1 mM $MgSO_4$, pH 8.3, 1 $\mathrm{g\,l^{-1}}$ bovine serum albumin and 0.8% Antibiotic Antimycotic Solution) for surgery. Using Dumostar-Biology 55 forceps (Dumont), the vitelline membrane was removed and the animal cap was isolated from the embryo. The caps were finally transferred to a new dish or a chamber containing fresh DFA medium for imaging.

mRNA, embryo microinjection, and animal cap confocal microscopy. Plasmids for expression of EB3–GFP and histone H2B–RFP mRNAs were obtained at the

2013 Advanced Imaging in *Xenopus* Workshop from the Wallingford laboratory. The mRNAs were synthetized using an mMessage mMachine SP6 Transcription Kit (Ambion, Thermo Fisher Scientific) following the supplier's protocol. The mRNAs were purified using phenol–chloroform extraction, resuspended in ddH$_2$O, aliquoted, and stored at −80 °C.

At stage 2, $t_e \times l_s$ hybrid embryos were transferred to 1/9× MMR 3% Ficoll. A solution containing 50 pg nl^{-1} of H2B–RFP mRNA and 100 pg nl^{-1} of EB3–GFP mRNA, concentrations that allowed us to image fluorescent signal as early as stage 9, was loaded into a needle pulled from a 1-mm glass capillary tube (TW100F-4, World Precision Instruments) using a P-87 Micropipette Puller (Sutter Instrument). Embryos were placed in a mesh-bottomed dish and microinjected in both blastomeres with 1 nl of the mRNA solution using a Picospritzer III microinjection system (Parker) equipped with a MM-3 micromanipulator (Narishige). Injected embryos were transferred to a new dish and incubated at 23 °C in 1/9× MMR 3% Ficoll until stage 8, when they were processed for animal cap isolation as described above. Caps were placed in a chamber filled with DFA medium made using 1 cm × 1 cm Gene Frames (Thermo Fisher Scientific) between a slide and a coverslip (Thermo Fisher Scientific) for confocal microscopy.

Embryo whole-mount immunofluorescence. At desired stages, embryos were fixed for 1–3 h using either MAD fixative (two parts of methanol (Thermo Fisher Scientific), two parts of acetone (Thermo Fisher Scientific), one part of DMSO) for most antibodies or MEMFA fixative (0.1 M MOPS pH 7.4, 2 mM EGTA, 1 mM MgSO$_4$, 3.7% formaldehyde) for the γH2A.X antibody. After fixation, embryos were dehydrated in methanol and stored at −20 °C. Embryos were then processed as previously described[28] with some modifications. Following gradual rehydration in 0.5× SSC (1× SSC: 150 mM NaCl, 15 mM Na citrate, pH 7.0), embryos were bleached with 1–2% H$_2$O$_2$ (Thermo Fisher Scientific) in 0.5× SSC containing 5% formamide for 2–3 h under light, then washed in PBT, a PBS solution containing 0.1% Triton X-100 (Thermo Fisher Scientific), and 2 mg ml^{-1} bovine serum albumin. Embryos were blocked in PBT supplemented with 10% goat serum (Gibco, Thermo Fisher Scientific) and 5% DMSO for 1–3 h and incubated overnight at 4 °C in PBT supplemented with 10% goat serum and the primary antibodies. We used different combinations of the following antibodies: 1:500 mouse anti-β-tubulin (E7; Developmental Studies Hybridoma Bank), 1:500 rabbit anti-histone H3 (ab1791; Abcam), 1:500 rabbit anti-lamin B1 (ab16048; Abcam), and 1:500 mouse anti-phospho-histone H2A.X (05-636; EMD Millipore, Merck KGaA). Embryos were then washed 4 × 2 h in PBT and incubated overnight in PBT supplemented with 1:500 goat anti-mouse or goat anti-rabbit secondary antibodies coupled either to Alexa Fluor 488 or 568 (Invitrogen, Thermo Fisher Scientific) and with 1:200 YO-PRO iodide (Thermo Fisher Scientific) if the use of anti-histone H3 antibody as primary was not possible. Embryos were then washed 4 × 2 h in PBT and gradually dehydrated in methanol. Embryos were finally cleared in Murray's clearing medium (two parts of benzyl benzoate, one part of benzyl alcohol). Embryos were placed either in a chamber made using a flat nylon washer (Grainger) attached with nail polish (Sally Hansen) to a slide and covered by a coverslip or a chamber made of silicon grease (Beckman coulter) between slide and coverslip, and filled with Murray's clearing medium for confocal microscopy.

Confocal microscopy, micronuclei, and nuclear size quantification. Confocal microscopy was performed on a Zeiss LSM 780 NLO AxioExaminer using the Zeiss Zen software. For animal cap live imaging, histone H2B–RFP and EB3–GFP signals were imaged on a single plane with a frame size of 1,024 pixels × 1,024 pixels every 5 s using a Plan-Apochromat 40×/1.4 oil objective and laser power of 22%. For imaging of histone H3, embryos were imaged using a Plan-Apochromat 20×/1.0 water objective and laser power of 12%, on multiple 1,024 × 1,024 pixel plans spaced of 0.68 μm in Z. For characterization of the micronuclei (lamin B1 and γH2A-X), embryos were imaged using a Plan-Apochromat 63×/1.40 oil objective and laser power of 12%, on multiple plans spaced 0.38 μm in Z. Images are mean averages of two scans with a depth of 16 bits. Pinhole size was always chosen to correspond to 1 Airy unit.

Micronuclei were quantified at stages 4, 6, 7, 8, and 9 as the number of observed micronuclei in the dataset divided by the number of nuclei in the dataset. The number of micronuclei at all stages and of nuclei at stages 4 and 6 were counted manually in Fiji. The number of nuclei at stages 7, 8, and 9 was determined automatically through histone H3 fluorescence signal segmentation using Imaris (Bitplane). Nuclear area in $tt_e \times t_s$, *X. tropicalis* and $t_e \times [t_s]$ was measured in Fiji using the ellipse tool. From this, we calculated the diameter of a circle of the same area, a value that we could directly compare to the cell size determined through the measurement of the cell diameter at the nucleus central plan.

Embryo nuclei purification. Embryo nuclei were prepared as previously described[29] from *X. tropicalis*, $l_e \times t_s$ hybrid, and *X. laevis* embryos. In brief, embryos were arrested at stage 8 in late interphase using 150 μg ml^{-1} cycloheximide in 1/10× MMR for 60 min. Then they were washed several times in ELB (250 mM sucrose, 50 mM KCl, 2.5 mM MgCl$_2$, and 10 mM HEPES pH 7.8) supplemented with LPC (10 μg ml^{-1} each leupeptin, pepstatin, chymostatin), cytochalasin D (100 μg ml^{-1}), and cycloheximide (100 μg ml^{-1}), packed in a tabletop centrifuge at 200*g* for 1 min, crushed with a pestle, and centrifuged at 10,000*g* for 10 min at 16 °C. The cytoplasmic extract containing endogenous embryonic nuclei was collected, supplemented with 8% glycerol, aliquoted, frozen in liquid nitrogen, and stored at −80 °C.

***Xenopus* egg extracts and related methods.** *X. laevis*[30] and *X. tropicalis*[4] metaphase-arrested egg extracts were prepared and spindle reactions conducted as previously described.

To reconstitute spindle assembly, stage 8 embryo nuclei were used as a source of DNA. Aliquots were thawed, resuspended in 1 ml of ELB, and spun at 1,600*g* for 5 min at room temperature. Pelleted nuclei were resuspended in 25 μl of fresh *X. tropicalis* extract and incubated at room temperature.

To examine kinetochore assembly, *X. laevis* sperm nuclei, prepared as previously described[31], were used as a source of DNA in both *X. laevis* and *X. tropicalis* egg extracts. Cycled chromosomes were prepared and spun-down[30], then processed for immunofluorescence as previously described[32]. In brief, the coverslips were incubated for 1 min in cold methanol, washed with PBS+NP40, and blocked overnight in PBS + 5% bovine serum albumin at 4 °C. The anti-Ndc80 (1:300 dilution, Stukenberg laboratory) or the anti-CENP-A (1:500 dilution, Straight laboratory) rabbit antibodies were added for 1 h. After washing with PBS+NP40, the coverslips were incubated with 1:1,000 anti-rabbit antibody coupled to Alexa Fluor 488 (Invitrogen, Thermo Fisher Scientific) for 30 min and then with 1:1,000 Hoechst for 5 min. The coverslips were finally washed and mounted for imaging. Each presented dataset was obtained from three different egg extracts with technical duplicates for each.

Spindles and chromosomes were imaged using micromanager software[33] with an Olympus BX51 microscope equipped with an ORCA-ER or an ORCA-II camera (Hamamatsu Photonics), and with an Olympus UPlan FL 40× air objective. Spindle measurements were made using Fiji and the spindle tubulin intensity line scan using an automated Java ImageJ plugin developed by X. Zhou (https://github.com/XiaoMutt/AiSpindle).

TUNEL assay. Embryos were fixed in MEMFA as described for embryo whole-mount immunofluorescence and processed as previously described[20] with minor modifications. In brief, after gradual rehydration, embryos were bleached with 1–2% H$_2$O$_2$ in 0.5× SSC containing 5% formamide for 1–2 h under light. After washes in PBS, embryos were incubated in 1× terminal deoxynucleotidyl transferase (TdT) buffer for 1 h and then overnight in TdT Buffer supplemented with 150 U ml^{-1} of TdT enzyme (Invitrogen, Thermo Fisher Scientific) and 1 pmol μl^{-1} Digoxigenin-11-dUTP (Roche). After washes in 1 mM EDTA/PBS at 65 °C, in PBS, and then in MAB (100 mM maleic acid, 150 mM NaCl, pH 7.5), embryos were blocked for 1 h in 2% Blocking Reagent (Roche) in MAB and then incubated overnight at 4 °C in 2% Blocking Reagent in MAB supplemented with 1:3,000 anti-Digoxigenin AP antibody (Roche). After washes in MAB and in AP Buffer (100 mM Tris, pH 9.5, 50 mM MgCl$_2$, 100 mM NaCl, 0.1% Tween 20, 2 mM levamisol), embryos were stained with NBT/BCIP (nitro-blue tetrazolium/5-bromo-4-chloro-3-indolyl phosphate; Promega) diluted in AP buffer. Reactions were stopped in MAB and embryos fixed overnight in Bouin's solution. After washes in 70% buffered ethanol and in methanol, embryos were imaged in methanol with the ToupView software (ToupTek) using an AmScope MD200 USB camera mounted on M5 stereoscope (Wild Heerbrugg).

Nucleic acid isolation, library construction, and sequencing. For genomic DNA, embryos at desired stages were incubated overnight in lysis buffer (50 mM Tris-HCl, 5 mM EDTA, 100 mM NaCl, 0.5% SDS) containing 250 μg ml^{-1} Proteinase K (Roche). DNA was isolated using phenol–chloroform extraction and ethanol precipitation. To isolate RNAs, embryos at desired stages were homogenize mechanically in TRIzol (Thermo Fisher Scientific) using up to a 30-gauge needle and processed according to the supplier's instructions. After resuspension in nuclease-free H$_2$O, RNAs were cleaned up using RNeasy kit (Qiagen) with on-column DNA digestion, following the supplier's protocol.

Libraries were constructed at the Functional Genomics Laboratory (FGL), a QB3-Berkeley Core Research Facility at UC Berkeley. For genomic DNA, an S220 Focused-Ultrasonicator (Covaris) was used to fragment DNA. The fragmented DNA was cleaned and concentrated with an MinElute PCR Purification kit (Qiagen). The library preparation was done on an Apollo 324 with PrepX ILM 32i DNA Library Kits (WaferGen Biosystems), and seven cycles of PCR amplification were used for library fragment enrichment. For RNAs, mRNA enrichment was performed on total RNA using polyA selection with an Invitrogen Dynabeads mRNA Direct kit. The library preparation was done on an Apollo324 with PrepX RNA-seq Library Prep Kits (WaferGen Biosystems), and 13 cycles of PCR amplification were used for index addition and library fragment enrichment.

Sequencing was performed by the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley. All samples were run as 100 paired-end HiSeq4000 lanes, pooled equimolar after quantification using KAPA Illumina Library quantification qPCR reagents on the BioRad CFX connect. Demultiplexing was performed to allow a single mismatch with Illumina's bcl2fastq version 2.17 software.

Genomic DNA sequencing analysis and deletion detection in hybrids. DNA sequencing reads were mapped to a *X. laevis*–*X. tropicalis* hybrid genome (Xenla9.1 and Xentro9.0) or a *X. borealis*–*X. tropicalis* hybrid genome (Xbo_04Apr2017 (A. Mudd and D. Rokhsar, unpublished observations) and Xentro9.0) using bwa mem (version 0.7.10-r789) with default settings. Duplicate reads were marked using bamUtil version 1.0.2.

Deletions in $t_e \times l_s$ and $l_e \times t_s$ hybrids were called by comparing local DNA sequencing read coverage between hybrid ($t_e \times l_s$ or $l_e \times t_s$) and parental ($t_e \times t_s$ or $l_e \times l_s$) genomes. The read coverage was determined in 10-kb regions, with a 2-kb sliding window across the genome. For each 10-kb region, we calculated the reads per kilobase per million reads (RPKM) in $t_e \times l_s$, $l_e \times t_s$, $t_e \times t_s$, and $l_e \times l_s$ sequencing tracks. The ratio of median RPKM values in retained regions had a non-zero baseline as expected because of a different size of hybrid and parental genomes. The ratio cut-off for deleted regions was set accordingly at fourfold and sixfold for *X. laevis* and *X. tropicalis* sequences, respectively. Lost regions overlapping for more than 30% of their length with gaps were removed and regions within 10 kb of each other were merged. Lost genes were analysed using the PANTHER database[24]. Because PANTHER only provided adjusted *P* values for fold enrichment based on binomial tests, we additionally estimated the 95% confidence interval for each enriched pathway using binom.test() function in R (version 3.2.2).

Deletions in $t_e \times b_s$ hybrids were called by identifying reduced genomic DNA signal in $t_e \times b_s$. The RPKM read coverage was determined in 10-kb regions, with a 2-kb sliding window across the genome. Regions with median $\log_{10}$ RPKM less than -1.25 in the *X. tropicalis* genome and -1.15 in the *X. borealis* genome were marked as deleted. Lost regions overlapping for more than 30% of their length with gaps were removed and regions within 10 kb of each other were merged.

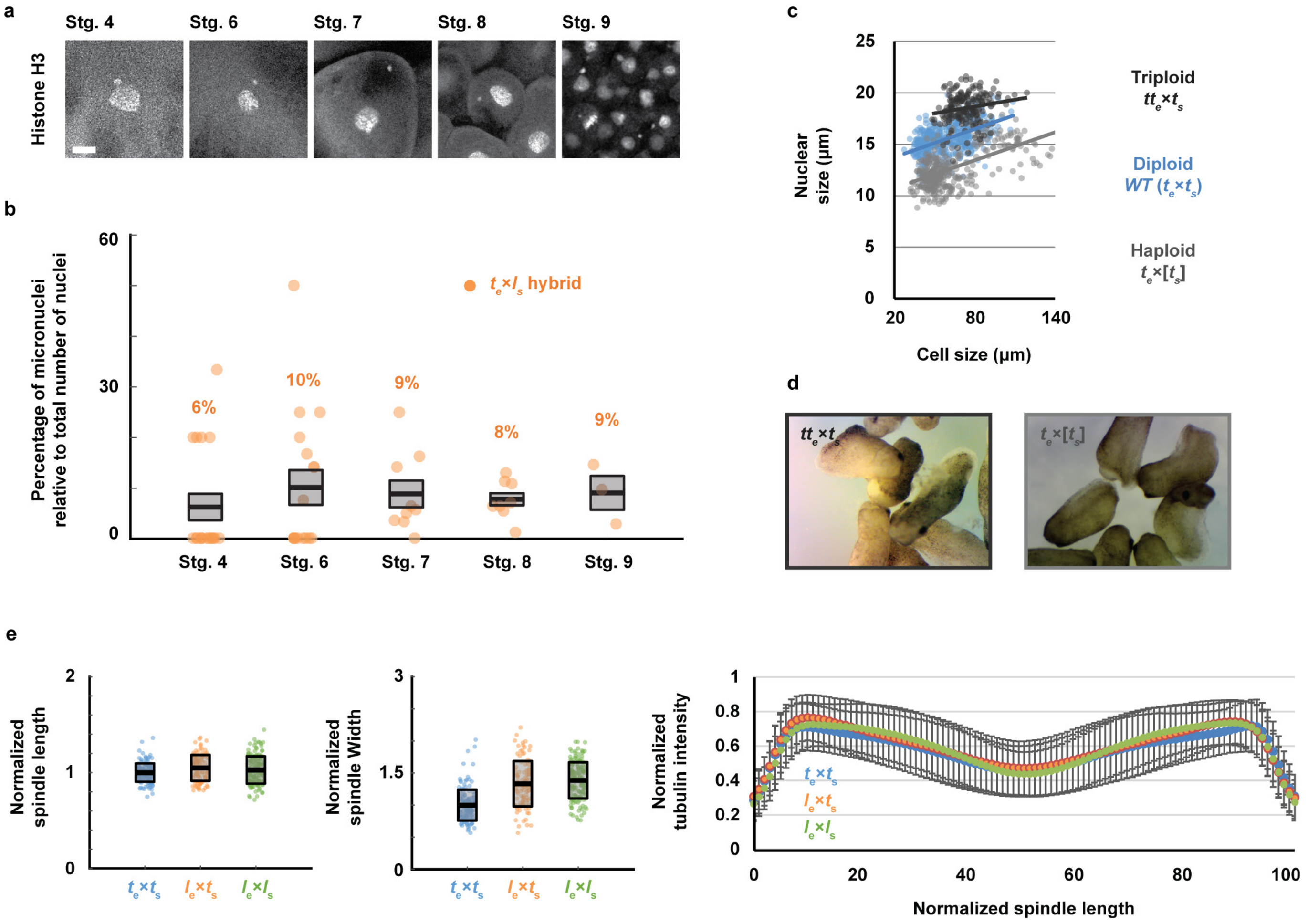
RNA sequencing analysis. We mapped RNA sequencing reads to the combined primary transcripts of *X. laevis* (JGIv18pV4) and *X. tropicalis* (JGIv91) using bwa mem (version 0.7.10), and discarded all reads mapped in multiple targets for further analysis. To use human gene annotation, which is more comprehensive than *Xenopus*, we transferred the expression level of these species to human orthologues. On the basis of the best BLASTP hit to human longest protein sequences (from EnsEMBL version 80), we merged the read counts of *X. laevis* and *X. tropicalis* genes to corresponding human genes, then performed differential expression analysis with EdgeR[34]. An adjusted *P* value criterion of less than 0.05 was applied to determine the significance of differential expression. To estimate the 95% confidence interval of log-scale fold change, we used limma[35] (version 3.28.10). Results of both EdgeR and Limma analyses are presented in Supplementary Table 2. For metabolic gene analysis, we used the list of metabolic genes obtained from the PANTHER database[24].

Metabolomic profiling. Seven hours post-fertilization, $t_e \times l_s$ hybrid, $t_e \times b_s$ hybrid, and respective *X. tropicalis* control embryos were collected from three independent fertilizations, always using eggs from the same female between the $t_e \times l_s$ hybrid or the $t_e \times b_s$ hybrid, and its *X. tropicalis* control. Five samples of 8 embryos each for non-polar lipid metabolites and 12 embryos each for polar metabolites were rinsed twice in filtered PBS and frozen in liquid nitrogen. Non-polar lipid metabolites from the eight embryos were extracted in 3 ml of 2:1 chloroform:methanol and 1 ml of PBS with inclusion of internal standards C12 monoalkylglycerol ether (10 nmol, Santa Cruz Biotechnology) and pentadecanoic acid (10 nmol). Organic and aqueous layers were separated by centrifugation at 1,000*g* for 5 min and the organic layer was collected, dried under a stream of nitrogen, and dissolved in 120 μl chloroform. Polar metabolites were extracted from the 12 embryos in 180 μl of 40:40:20 (acetonitrile:MeOH:H_2O) with inclusion of internal standard D3N15 serine (50 nM, Cambridge Isotope Laboratories, DNLM-6863). Samples were disrupted by sonication then centrifuged at 21,000*g* for 10 min and the supernatant was collected for analysis. Metabolites were separated by liquid chromatography, and MS analysis was performed with an electrospray ionization source on an Agilent 6430 QQQ LC-MS/MS (Agilent Technologies). The capillary voltage was set to 3.0 kV, and the fragmentor voltage was set to 100 V. The drying gas temperature was 350 °C, the drying gas flow rate was 10 l min^{-1}, and the nebulizer pressure was 35 p.s.i. Metabolites were identified by single-reaction monitoring of the transition from precursor to product ions at associated optimized collision energies and retention times as previously described[36]. Metabolites were quantified by integrating the area under the curve, then normalized to internal standard values and tissue weight. Metabolite levels are expressed as relative abundances compared with controls. *P* values were calculated using a two-tailed homoscedastic *t*-test. Significance was analysed for an $\alpha = 0.05$ threshold as well as that with a Bonferroni-like correction to account for multiple hypothesis comparison. Strict Bonferroni correction is highly conservative and often results in increased type II errors (failing to acknowledge a real effect) and several alternatives exist. Because we compared around 200 compounds between two types of embryo, each with five replicates, we used a penalized Bonferroni correction, and divided the α threshold value by the logarithm of the number of tests ($\alpha/\log(k)$) to decrease the risk of a type II error.

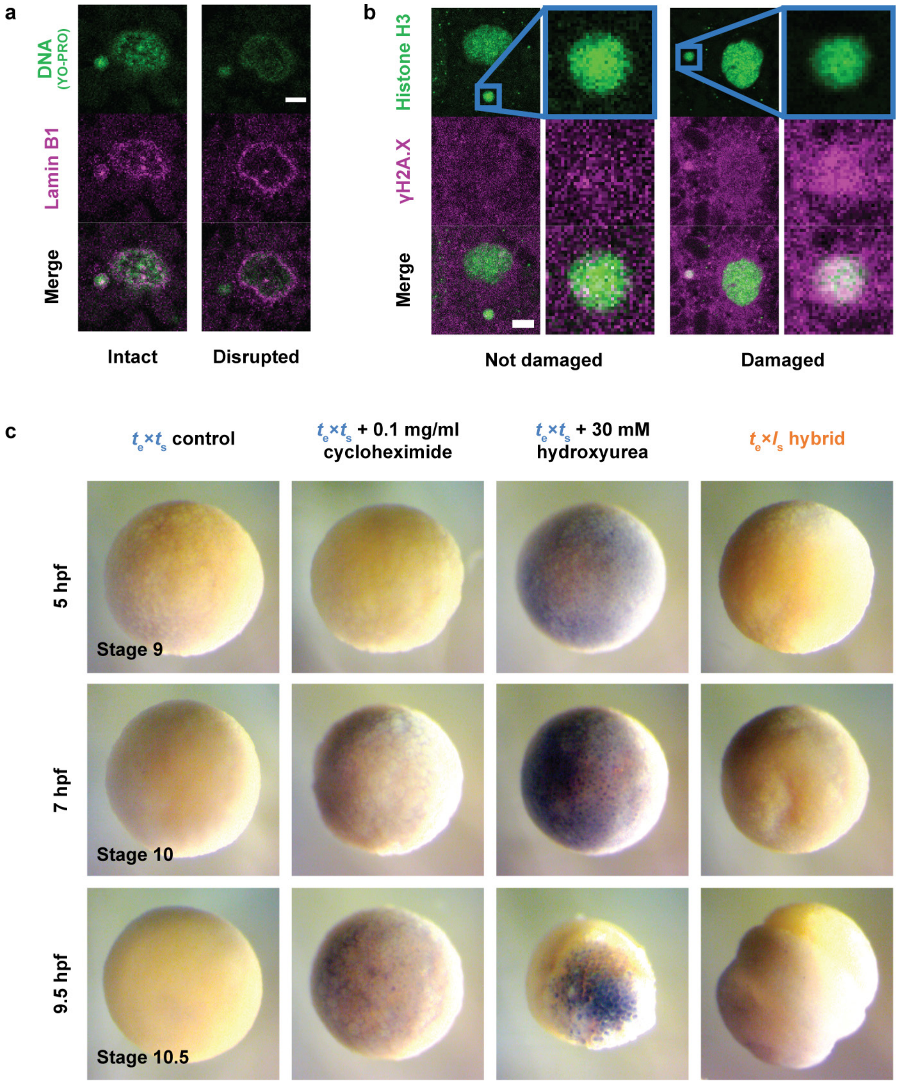
Data availability. All genomic and transcriptomic data generated for this study are available from public databases: stage 9 $t_e \times l_s$ hybrid whole-genome sequencing data (NCBI Sequence Read Archive SRP124316) and corresponding stage 9 *X. laevis*, *X. tropicalis*, and $l_e \times t_s$ hybrid controls (NCBI Gene Expression Omnibus GSE92382), tailbud and tadpole stage $tt_e \times l_s$ hybrid whole-genome sequencing data (NCBI Sequence Read Archive SRP124316), stage 9 $t_e \times b_s$ hybrid whole-genome sequencing data (NCBI Sequence Read Archive SRP124316), and 7 h.p.f. *X. tropicalis* and $t_e \times l_s$ hybrid transcriptome RNA sequencing data (NCBI Gene Expression Omnibus GSE106157).

Code availability. MATLAB scripts were written to acquire and process embryo live-imaging movies. All these scripts are available from the corresponding author upon request. Two MATLAB scripts used here for movie cropping and concatenation were obtained through the MATLAB Central File Exchange: 'videoMultiCrop' and 'concatVideo2D' by 'Nikolay S'. The automated spindle tubulin intensity line scan Java ImageJ plugin developed by X. Zhou is available on GitHub (https://github.com/XiaoMutt/AiSpindle).

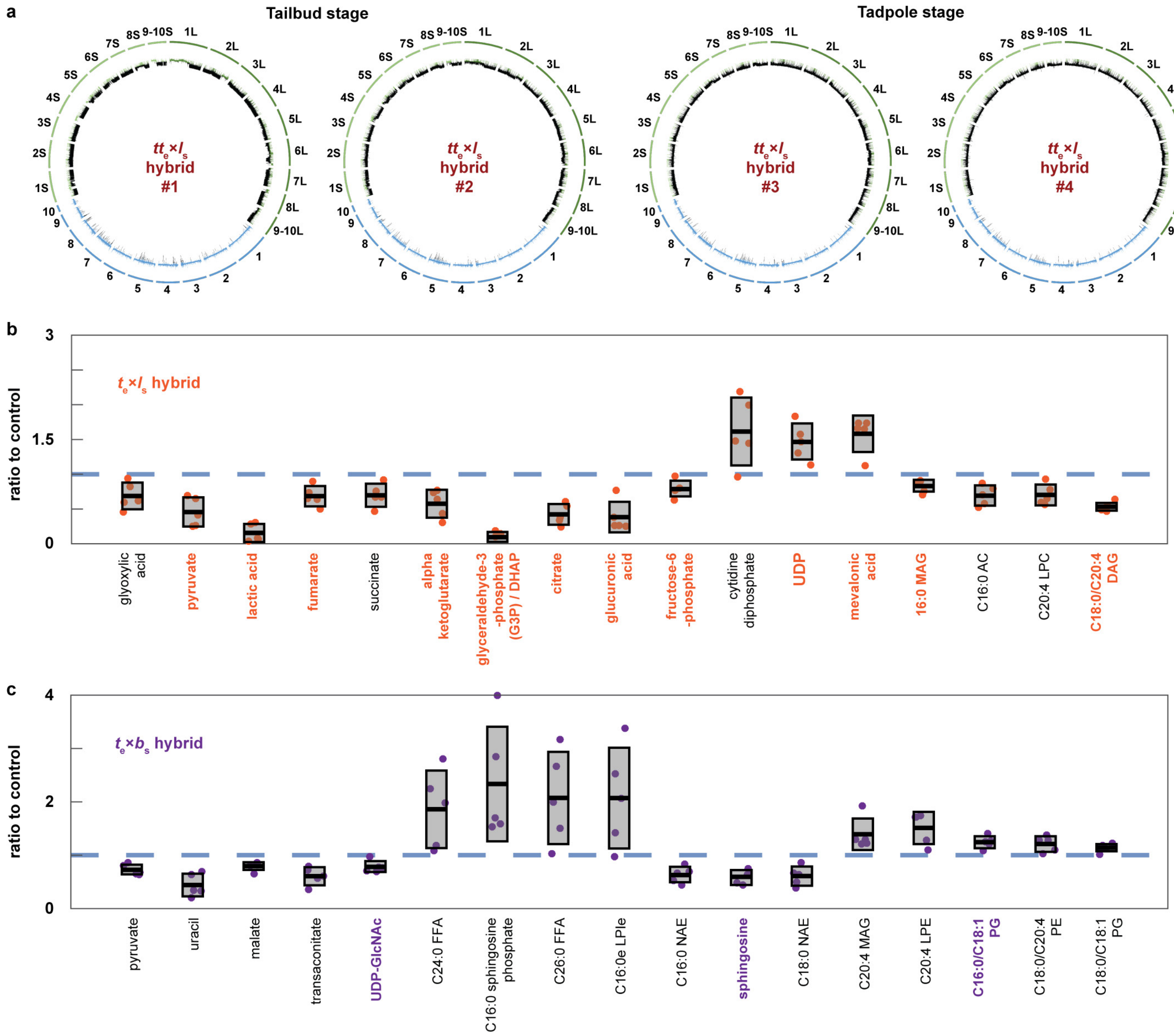
26. Nieuwkoop, P. D & Faber, J. *Normal Table of Xenopus laevis (Daudin)* (Garland, 1994).
27. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9,** 676–682 (2012).
28. Lee, C., Kieserman, E., Gray, R. S., Park, T. J. & Wallingford, J. Whole-mount fluorescence immunocytochemistry on *Xenopus* embryos. *CSH Protoc.* **2008,** pdb.prot4957 (2008).
29. Levy, D. L. & Heald, R. Nuclear size is regulated by importin α and Ntf2 in *Xenopus*. *Cell* **143,** 288–298 (2010).
30. Maresca, T. J. & Heald, R. Methods for studying spindle assembly and chromosome condensation in *Xenopus* egg extracts. *Methods Mol. Biol.* **322,** 459–474 (2006).
31. Murray, A. W. Cell cycle extracts. *Methods Cell Biol.* **36,** 581–605 (1991).
32. Hannak, E. & Heald, R. Investigating mitotic spindle assembly and function *in vitro* using *Xenopus* laevis egg extracts. *Nat. Protocols* **1,** 2305–2314 (2006).
33. Edelstein, A. D. *et al.* Advanced methods of microscope control using μManager software. *J. Biol. Methods* **1,** 10 (2014).
34. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
35. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
36. Louie, S. M. *et al.* GSTP1 is a driver of triple-negative breast cancer cell metabolism and pathogenicity. *Cell Chem. Biol.* **23,** 567–578 (2016).



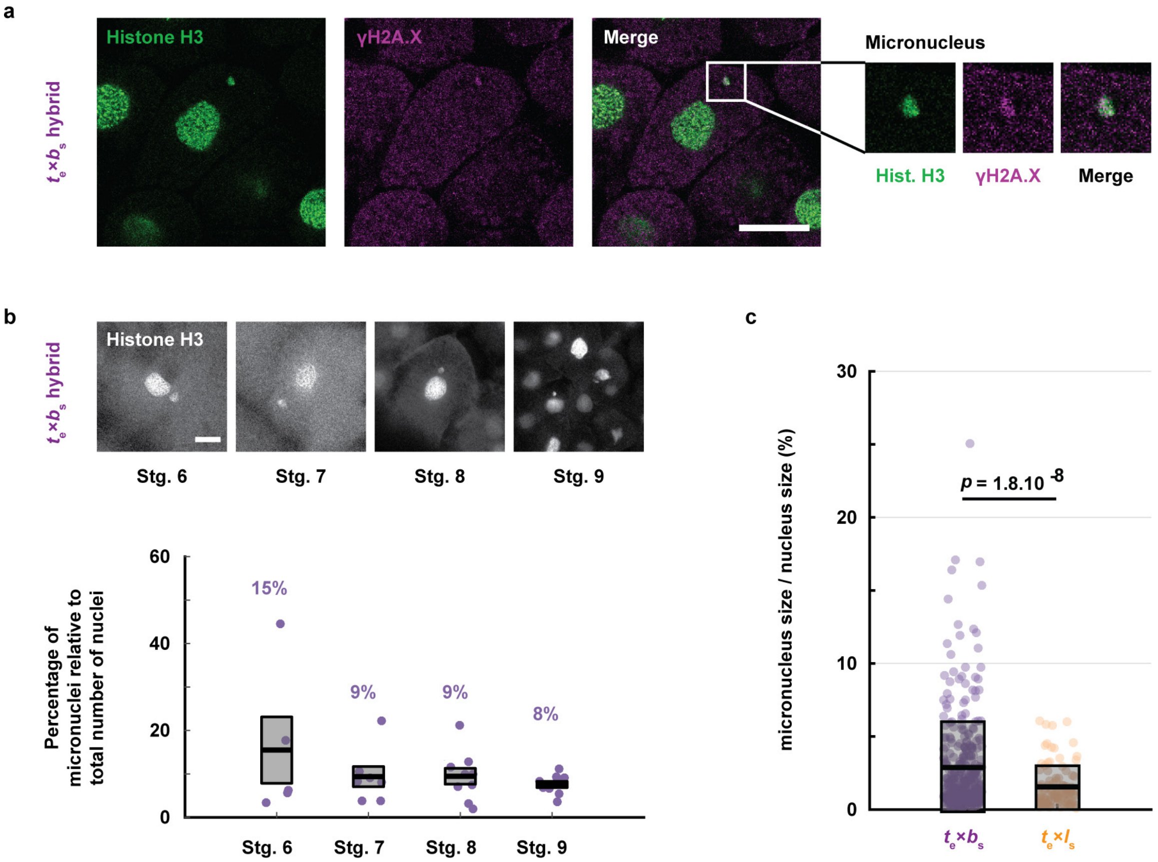
Extended Data Figure 1 | Occurrence of micronuclei, role of ploidy and spindle architecture. a, Micronuclei in $t_e \times l_s$ hybrid embryos at various developmental stages. Whole-mount embryo immunofluorescence was performed in $t_e \times l_s$ hybrid embryos using anti-histone H3 antibody at stages 4, 6, 7, 8, and 9 and quantified in **b**. Scale bar, 10 μm. **b**, Quantification of micronuclei in $t_e \times l_s$ hybrid embryos. The percentage of micronuclei was calculated as the number of micronuclei in the imaged portion of the embryo divided by the total number of nuclei in the same imaged portion. The average percentage for multiple embryos at stage 4 ($n = 18$ $t_e \times l_s$ hybrid embryos (individual dots) with a total of 63 nuclei), stage 6 ($n = 17/115$), stage 7 ($n = 9/322$), stage 8 ($n = 8/1119$), and stage 9 ($n = 3/2004$) from three independent experiments is shown as thick line. Grey boxes indicate 1 s.e.m. Control *X. tropicalis* embryos from the same mothers were analysed but no micronuclei were observed at any stages. **c**, Nuclear size in *X. tropicalis* embryos with varying ploidy. Nuclear size relative to cell size (diameters in micrometres) is plotted for triploid ($tt_e \times t_s$; dark grey, $n = 175$ nuclei from six embryos), diploid (*X. tropicalis*, $t_e \times t_s$; blue, $n = 453/9$), and haploid ($t_e \times [t_s]$; light grey, $n = 346/16$) embryos. Each dot indicates an individual data point and the solid lines indicate a linear fit. **d**, *X. tropicalis* embryos with varying ploidy at tailbud stage. Images of triploid ($tt_e \times t_s$; left) and haploid ($t_e \times [t_s]$; right) tailbuds were taken under identical conditions. Similar observations were over three independent experiments. **e**, Size and microtubule distribution in *X. tropicalis* spindles assembled from different embryo nuclei DNA ($n = 147$, 103, and 156 spindles quantified for *X. tropicalis*, $l_e \times t_s$ hybrids, and *X. laevis* embryo nuclei, respectively, from three different egg extracts). Spindle length (left) and width (middle) were normalized to the *X. tropicalis* control, averages are shown as thick black lines, and the grey boxes indicate 1 s.d. Ninety-five per cent confidence intervals for lengths are 1 ± 0.02 for $t_e \times t_s$, 1.05 ± 0.03 for $l_e \times t_s$, and 1.03 ± 0.02 for $l_e \times l_s$, and for widths are 1 ± 0.04, 1.3 ± 0.07, and 1.4 ± 0.04. Line scans of rhodamine-tubulin signal along spindle length were taken (right). Spindle lengths were normalized to 100% and tubulin intensities were normalized within datasets. The average intensities are plotted for the three spindle types, error bars indicate s.d., and colours are as in Fig. 2a.



Extended Data Figure 2 | Characterization of micronuclei in $t_e \times l_s$ hybrid embryos and link to embryo death. a, Disrupted micronuclei envelopes in $t_e \times l_s$ hybrid embryos. Whole-mount embryo immunofluorescence was performed in $t_e \times l_s$ hybrid embryos using the YO-PRO DNA dye (top) and anti-Lamin B1 antibody (middle); corresponding channels are shown in green and magenta, respectively. The merged images are shown below. Twenty-five micronuclei within five different embryos were analysed. Intact (left) and disrupted (right) envelopes were observed in all analysed embryos. Scale bar, 10 μm. **b**, DNA damage in $t_e \times l_s$ hybrid embryo micronuclei. Whole-mount embryo immunofluorescence was performed in $t_e \times l_s$ hybrid embryos using anti-histone H3 (top) and anti-γH2A.X (middle) antibodies; corresponding channels are shown in green and magenta, respectively. The merge images are shown below. Twenty-one micronuclei within different six embryos were analysed. Micronuclei with undamaged (left; negative γH2A.X signal) and damaged (right; positive γH2A.X signal) DNA were observed in all analysed embryos. Zoomed images of micronuclei are shown on the right of each image. Scale bar, 10 μm. **c**, TUNEL assay in apoptotic *X. tropicalis* and $t_e \times l_s$ hybrid embryos. *X. tropicalis* (left), *X. tropicalis* treated with cycloheximide (middle left) or hydroxyurea (middle right) as indicated, and $t_e \times l_s$ hybrid (right) embryos were prepared for TUNEL assay 5 h.p.f. (equivalent stage 9; top), 7 h.p.f. (equivalent stage 10; middle), and 9.5 h.p.f. (equivalent stage 10.5; bottom). Identical results were obtained over three different experiments. Representative images are shown and were taken under identical conditions.



Extended Data Figure 3 | Whole-genome sequencing of $tt_e \times l_s$ rescued embryos and metabolomic profiling of $t_e \times l_s$ and $t_e \times b_s$ hybrid embryos. a, The genomes of 4 $tt_e \times l_s$ rescued embryos were sequenced, aligned, and normalized to the genomes of *X. tropicalis* (blue) and *X. laevis* (green) for which sub-genomes S and L were distinguished (S in light green and L in dark green). Underrepresented regions of the genomes are colour-coded in black. The $tt_e \times l_s$ embryo genomes 1 and 2 were prepared from tailbuds, and 3 and 4 from tadpoles. **b**, Metabolites differentially represented between $t_e \times l_s$ hybrid and *X. tropicalis* embryos 7h.p.f. Among the 179 metabolites detected, 17 were significantly altered in $t_e \times l_s$ hybrid embryos ($P < 0.05$; two-tailed homoscedastic t-test; individual P values are provided in Fig. 3c source data) and are shown as a ratio to the *X. tropicalis* control (blue dashed line). Levels were obtained from five samples from three independent fertilizations each. Values for the $t_e \times l_s$ hybrid are plotted in orange. The averages are shown as thick lines and the grey boxes correspond to 1 s.d. Ninety-five per cent confidence intervals are, from left to right, 0.69 ± 0.24, 0.46 ± 0.26, 0.16 ± 0.16, 0.68 ± 0.18, 0.70 ± 0.21, 0.58 ± 0.25, 0.10 ± 0.09, 0.42 ± 0.19, 0.38 ± 0.27, 0.79 ± 0.15, 1.61 ± 0.61, 1.47 ± 0.33, 1.58 ± 0.33, 0.83 ± 0.11, 0.71 ± 0.18, 0.70 ± 0.19, and 0.53 ± 0.08. Metabolites with P values below the penalized Bonferroni corrected threshold ($n = 12$) are labelled in orange. **c**, Metabolites differentially represented between $t_e \times b_s$ hybrid and *X. tropicalis* embryos 7 h.p.f. Among the 241 metabolites detected, 17 were significantly altered in $t_e \times b_s$ hybrid embryos ($P < 0.05$; two-tailed homoscedastic t-test; individual P values are provided in Fig. 4g source data) and are shown as a ratio to the *X. tropicalis* control (blue dashed line). Levels were obtained from five samples from three independent fertilizations, each. Values for the $t_e \times b_s$ hybrid are plotted in purple. The averages are shown as thick lines and the grey boxes correspond to 1 s.d. Ninety-five per cent confidence intervals are, from left to right, 0.73 ± 0.12, 0.44 ± 0.26, 0.80 ± 0.10, 0.61 ± 0.21, 0.78 ± 0.14, 1.86 ± 0.9, 2.33 ± 1.33, 2.07 ± 1.07, 2.07 ± 1.17, 0.63 ± 0.19, 0.59 ± 0.16 0.61 ± 0.22, 1.39 ± 0.37, 1.51 ± 0.38, 1.24 ± 0.14, 1.21 ± 0.18, and 1.14 ± 0.10. Metabolites with P values below the penalized Bonferroni corrected threshold ($n = 3$) are labelled in purple.



Extended Data Figure 4 | Characterization of micronuclei in $t_e \times b_s$ hybrid embryos. a, DNA damage in $t_e \times b_s$ hybrid embryo micronuclei. Whole-mount embryo immunofluorescence was performed in $t_e \times b_s$ hybrid embryos using anti-histone H3 (left) and anti-γH2A.X (middle) antibodies; corresponding channels are shown in green and magenta, respectively. The merged image is shown on the right. Thirty-four micronuclei within eight different embryos were analysed. Micronuclei with damaged DNA were observed in all analysed embryos. Zoomed images of micronuclei are shown on the right in the same left-to-right order. Scale bar, 20 μm. **b**, Micronuclei in $t_e \times b_s$ hybrid embryos at various developmental stages (top). Whole-mount embryo immunofluorescence was performed in $t_e \times b_s$ hybrid embryos using anti-histone H3 antibody at stages 6, 7, 8, and 9. Scale bar, 20 μm. Quantification of micronuclei in $t_e \times b_s$ hybrid embryos (bottom). The percentage of micronuclei was calculated as the number of micronuclei in the imaged portion of the embryo divided by the total number of nuclei in the same imaged portion. The average percentage for multiple embryos at stage 6 ($n = 5$ $t_e \times b_s$ hybrid embryos (individual dots) with a total of 125 nuclei), stage 7 ($n = 7/153$), stage 8 ($n = 9/731$), and stage 9 ($n = 10/2,691$) is shown as a thick line. Grey boxes correspond to 1 s.e.m. Control *X. tropicalis* embryos from the same mothers were analysed but no micronuclei were observed at any stages. **c**, Micronuclei size in $t_e \times b_s$ and $t_e \times l_s$ hybrids. Size is plotted as the ratio between the volumes of the micronucleus and its corresponding nucleus. Each dot represents an individual data point ($n = 329$ micronuclei from 36 $t_e \times b_s$ embryos shown in purple and $n = 100$ from 17 $t_e \times l_s$ embryos shown in orange, from 4 independent experiments). The thick black line indicates the average and the grey box corresponds to 1 s.d. Ninety-five per cent confidence intervals are 2.9 ± 0.36% for $t_e \times b_s$ and 1.6 ± 0.28% for $t_e \times l_s$ embryos. Statistical significance was shown using a two-tailed heteroscedastic *t*-test.

Extended Data Table 1 | Embryonic development in *Xenopus* haploids and cybrids generated from *X. tropicalis* irradiated eggs

Embryo	Normal Stage 2[†] (n)	Regular Stage 9 (%)	Died between 9-13 (%)	Exogastrulae	Normal tailbuds	Abnormal tailbuds	Tadpoles	
							Stunted (%)	Normal (%)
$[t_e] \times t_s$	402 (5)	402 (100)	6 (1)	131 (33)	209 (52)	56 (14)	191 (48)	18 (4)
$[t_e] \times l_s$[‡]	25 (7)	25 (100)	25 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Symbol *n* indicates the number of different male–female combinations from which results were compiled.

†Unfertilized eggs and embryos that showed an abnormal or incomplete first cleavage were excluded from this analysis.

‡Fertilization efficiency of irradiated *X. tropicalis* eggs with *X. laevis* sperm was very low (~4%).

Extended Data Table 2 | Effects of drug treatments on $t_e \times t_s$ embryos

Drug	Inhibition	Time of addition	Concentration	Phenotype	Product details
Cycloheximide	Protein synthesis	stage 6.5	0.1 mg/ml	Cell cycle arrest at stage 7 followed by apoptosis	C7698 (Sigma-Aldrich)
Hydroxyurea	DNA replication	stage 3	30 mM	Apoptosis at late stage 8	AC151680050 (Thermo Fisher Sc.)
Triptolide	Transcription	stage 2	25 μM	Cell lysis at stage 9	T3652 (Sigma-Aldrich)
Olygomycin	ATP Synthase	stage 2	40 μM	Cell cycle arrest at stage 9	75351 (Sigma-Aldrich)
AP-III-a4	Enolase (including non-glycolytic functions)	stage 2	30 μM	Arrest at stage 7 and followed by cell lysis	19933 (Cayman Chemical)
Iodoacetic acid	Glyceraldehyde-3-P dehydrogenase	stage 2	50 mM	Cell lysis at stage 9	I4386 (Sigma-Aldrich)
CP-91,149	Glycogen phosphorylase	stage 2	270 μM	Cell death at stage 9 from the vegetal side	PZ0104 (Sigma-Aldrich)

X. tropicalis embryos were treated with different drugs at different stages. Phenotypes of effects are listed. When unspecified, apoptosis or lysis initiated at random locations in the embryo.

Extended Data Table 3 | Sub-genome distribution of lost compared with retained DNA in $t_e \times l_s$, $tt_e \times l_s$, and $t_e \times b_s$ hybrids

***t_e×l_s* hybrid**

Sub genome	Total (bp)	Lost (bp)	Remaining (bp)	Lost (%)	Remaining (%)
X. laevis L	1368982762	237294229	1131688533	17.33	82.67
X. laevis S	1139955720	11850000	1128105720	1.04	98.96
X. tropicalis	1272999256	3452000	1269547256	0.27	99.73

***tt_e×l_s* hybrid #1**

Sub genome	Total (bp)	Lost (bp)	Remaining (bp)	Lost (%)	Remaining (%)
X. laevis L	1368982762	1084660643	284322119	79.23	20.77
X. laevis S	1139955720	946048452	193907268	82.99	17.01
X. tropicalis	1272999256	5686000	1267313256	0.45	99.55

***tt_e×l_s* hybrid #2**

Sub genome	Total (bp)	Lost (bp)	Remaining (bp)	Lost (%)	Remaining (%)
X. laevis L	1368982762	1259268028	109714734	91.99	8.01
X. laevis S	1139955720	1003939197	136016523	88.07	11.93
X. tropicalis	1272999256	2964000	1270035256	0.23	99.77

***tt_e×l_s* hybrid #3**

Sub genome	Total (bp)	Lost (bp)	Remaining (bp)	Lost (%)	Remaining (%)
X. laevis L	1368982762	1360054762	8928000	99.35	0.65
X. laevis S	1139955720	1131571720	8384000	99.26	0.74
X. tropicalis	1272999256	3728000	1269271256	0.29	99.71

***tt_e×l_s* hybrid #4**

Sub genome	Total (bp)	Lost (bp)	Remaining (bp)	Lost (%)	Remaining (%)
X. laevis L	1368982762	1361764762	7218000	99.47	0.53
X. laevis S	1139955720	1134337720	5618000	99.51	0.49
X. tropicalis	1272999256	3240000	1269759256	0.25	99.75

***t_e×b_s* hybrid**

Sub genome	Total (bp)	Lost (bp)	Remaining (bp)	Lost (%)	Remaining (%)
X. borealis L	1428994000	108866000	1320128000	7.62%	92.38%
X. borealis S	1201786000	30804000	1170982000	2.56%	97.44%
X. tropicalis	1273010000	11592000	1261418000	0.91%	99.09%

Percentage of lost and remaining DNA for each sub-genome is shown for all hybrid genomes sequenced. Sub-genomes are colour-coded as in Figs 2c and 4e.

Extended Data Table 4 | Overrepresentation test of all or metabolism-only 3L and 4L lost genes

3L and 4L lost genes overrepresentation test	
Analysis Type	PANTHER Overrepresentation Test (release 20170413)
Annotation Version and Release Date	PANTHER version 11.1 Released 2016-10-24
Analyzed List	Client Text Box Input (Xenopus tropicalis)
Reference List	Xenopus tropicalis (all genes in database)
Bonferroni correction	TRUE

PANTHER GO-Slim Biological Process*	**Xenopus tropicalis - REFLIST (18238)**	**Client Text Box Input (843)**	**Client Text Box Input (expected)**	**Client Text Box Input (fold Enrichment)**	**Client Text Box Input (P-value)**	**95% Confidence Interval (binomial test)**
biosynthetic process (GO:0009058)	1295	141	100.05	1.41	8.01E-03	[123.4, ∞]
nitrogen compound metabolic process (GO:0006807)	1738	179	134.27	1.33	1.40E-02	[159.6, ∞]
metabolic process (GO:0008152)	6036	546	466.32	1.17	1.13E-03	[522.4, ∞]

3L and 4L lost metabolism genes overrepresentation test	
Analysis Type	PANTHER Overrepresentation Test (release 20170413)
Annotation Version and Release Date	PANTHER version 11.1 Released 2016-10-24
Analyzed List	Client Text Box Input (Xenopus tropicalis)
Reference List	Xenopus tropicalis (all genes in database)
Bonferroni correction	TRUE

PANTHER GO-Slim Biological Process†	**Xenopus tropicalis - REFLIST (18238)**	**Client Text Box Input (843)**	**Client Text Box Input (expected)**	**Client Text Box Input (fold Enrichment)**	**Client Text Box Input (P-value)**	**95% Confidence Interval (binomial test)**
glycolysis (GO:0006096)	26	6	0.78	7.71	3.71E-02	[2.6, ∞]
rRNA metabolic process (GO:0016072)	104	15	3.11	4.82	2.22E-04	[9.3, ∞]
DNA replication (GO:0006260)	114	16	3.41	4.69	1.38E-04	[10.1, ∞]
tRNA metabolic process (GO:0006399)	104	14	3.11	4.5	1.11E-03	[8.5, ∞]
generation of precursor metabolites and energy (GO:0006091)	185	24	5.54	4.33	9.87E-07	[16.6, ∞]

PANTHER software (http://pantherdb.org/) was used to perform a statistical overrepresentation test on all (top table) or metabolism-only (bottom table) lost genes from chromosomes 3L and 4L.
*Only over-represented processes are shown in the top table.
†Only the top five processes based on fold enrichment are shown in the bottom table.

doi:10.1038/nature25184

An extracellular network of *Arabidopsis* leucine-rich repeat receptor kinases

Elwira Smakowska-Luzan[1]*, G. Adam Mott[2]*, Katarzyna Parys[1]*, Martin Stegmann[3], Timothy C Howton[4], Mehdi Layeghifard[2], Jana Neuhold[5], Anita Lehner[5], Jixiang Kong[1], Karin Grünwald[1], Natascha Weinberger[1], Santosh B. Satbhai[1,6], Dominik Mayer[1,7,8], Wolfgang Busch[1,6], Mathias Madalinski[1,7,8], Peggy Stolt-Bergner[5], Nicholas J. Provart[2,9], M. Shahid Mukhtar[4], Cyril Zipfel[3], Darrell Desveaux[2,9], David S. Guttman[2,9] & Youssef Belkhadir[1]

The cells of multicellular organisms receive extracellular signals using surface receptors. The extracellular domains (ECDs) of cell surface receptors function as interaction platforms, and as regulatory modules of receptor activation[1,2]. Understanding how interactions between ECDs produce signal-competent receptor complexes is challenging because of their low biochemical tractability[3,4]. In plants, the discovery of ECD interactions is complicated by the massive expansion of receptor families, which creates tremendous potential for changeover in receptor interactions[5]. The largest of these families in *Arabidopsis thaliana* consists of 225 evolutionarily related leucine-rich repeat receptor kinases (LRR-RKs)[5], which function in the sensing of microorganisms, cell expansion, stomata development and stem-cell maintenance[6–9]. Although the principles that govern LRR-RK signalling activation are emerging[1,10], the systems-level organization of this family of proteins is unknown. Here, to address this, we investigated 40,000 potential ECD interactions using a sensitized high-throughput interaction assay[3], and produced an LRR-based cell surface interaction network ($\mathbf{CSI^{LRR}}$) that consists of 567 interactions. To demonstrate the power of $\mathbf{CSI^{LRR}}$ for detecting biologically relevant interactions, we predicted and validated the functions of uncharacterized LRR-RKs in plant growth and immunity. In addition, we show that $\mathbf{CSI^{LRR}}$ operates as a unified regulatory network in which the LRR-RKs most crucial for its overall structure are required to prevent the aberrant signalling of receptors that are several network-steps away. Thus, plants have evolved LRR-RK networks to process extracellular signals into carefully balanced responses.

LRR-RKs are modular proteins that feature an ECD with numerous LRR repeats, a transmembrane domain and an intracellular kinase domain[1]. LRR-RKs sense a wide array of endogenous and exogenous ligands, including peptides and small molecule hormones, to regulate development and immunity in plants[7,10]. Stereotypical LRR-RKs include the steroid receptor BRASSINOSTEROID INSENSITIVE1 (BRI1) as well as the immune receptors FLAGELLIN SENSING 2 (FLS2) and PLANT ELICITOR PEPTIDE RECEPTORS 1/2 (PEPR1/2)[1,11]. Ligand-induced activation of BRI1, FLS2 or PEPR1 and PEPR2 signalling requires physical interaction with the LRR-RK co-receptor BRI1-ASSOCIATED KINASE 1 (BAK1)[12–15]. In heterotypic LRR-RKs complexes, interactions between ECDs can activate or repress signalling pathways[2]. Yet, the full range of these interactions remains unmapped.

We cloned the ECDs of 200 LRR-RKs from *Arabidopsis* into bait and prey expression vectors for recombinant protein production in *Drosophila* Schneider S2 cells (Extended Data Fig. 1, Supplementary Table 1). We then implemented the extracellular interaction assay established previously[3] and performed an all-by-all screen of the 200 ECDs (Extended Data Fig. 2). Because the *Arabidopsis* genome encodes 225 LRR-RKs[5], we interrogated the extracellular LRRs interaction space to a completeness of 79%. This screen resulted in a $\mathrm{CSI^{LRR}}$ map containing 2,145 bidirectional interactions, of which only 26.4% (567 high-confidence interactions (HCI)) passed our extremely stringent statistical cut-offs for network construction (Fig. 1a, Supplementary Text 1, Supplementary Table 2). To verify our screen results, the ECDs from the 567 HCI and from a random set of 248 low-confidence interactions ($\mathrm{LCI^{CSI}}$) were independently re-expressed and retested. To benchmark the retest screen, we assembled a positive reference set of 20 literature-curated LRR-RK interaction pairs that complied with the criteria defined previously[4,16] (Supplementary Table 3). In the retest, the positive reference set, HCI and $\mathrm{LCI^{CSI}}$ scored positively at a rate of 100%, 92% and 12.5%, respectively (Extended Data Fig. 3, Supplementary Text 2, Supplementary Table 4). As expected for a high-quality set, the confirmation rates of the HCI and the positive reference set are statistically indistinguishable ($P = 0.3894$, two-tailed Fisher's exact test).

Models for LRR-RK signalling suggest that ECD interactions help to bring together the intracellular domains (ICDs) for subsequent interaction and signal transduction[2]. We therefore tested whether ICDs from 372 HCI were more likely to interact than another set of 50 randomly selected LRR-RKs via yeast two-hybrid (Y2H) assays ($\mathrm{LCI^{Y2H}}$)[17]. The HCI and $\mathrm{LCI^{Y2H}}$ scored positively at a rate of 54.3% and 10%, respectively (Supplementary Table 5). Notably, of the ICD interactions analysed by Y2H assays, ten were present in our positive reference set, and all tested positively (Supplementary Table 3, Supplementary Text 3). We assign an extremely high level of confidence to interactions that occur at both the ECD and ICD level.

Next, we investigated the biological relevance of $\mathrm{CSI^{LRR}}$ interactions by studying the ligand-dependent activation of BRI1- and FLS2-mediated signalling[1]. We compiled a collection of 27 transfer DNA (T-DNA) insertion mutants[18], targeting the HCI and LCI partners for both BRI1 and FLS2 ($\mathrm{HCI^{BRI1/FLS2}/LCI^{BRI1/FLS2}}$) (Extended Data Figs 4, 5, Supplementary Tables 6, 7). For these T-DNA lines, we used brassinosteroid-induced hypocotyl elongation assays to measure BRI1 activation, and bacterial flagellin peptide (flg22)-induced seedling growth inhibition, peroxidase activity and luminol-based reactive oxygen species (ROS) assays to measure FLS2 activation[19,20]. Although mutants corresponding to $\mathrm{HCI^{BRI1/FLS2}}$ partners showed

[1]Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna Biocenter (VBC), Dr Bohr-Gasse 3, 1030 Vienna, Austria. [2]Department of Cell & Systems Biology, University of Toronto, 25 Willcocks St., Toronto, Ontario, Canada. [3]The Sainsbury Laboratory, Norwich Research Park, Norwich NR4 7UH, UK. [4]Department of Biology, University of Alabama at Birmingham, Birmingham, Alabama, USA. [5]Protein Technologies Facility, Vienna Biocenter Core Facilities (VBCF), Vienna, Austria. [6]Salk Institute for Biological Studies, Plant Molecular and Cellular Biology Laboratory, 10010 N Torrey Pines Rd, La Jolla, California 92037, USA. [7]Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Campus-Vienna-Biocenter 1, 1030 Vienna, Austria. [8]Institute of Molecular Biotechnology GmbH (IMBA), Vienna Biocenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria. [9]Centre for the Analysis of Genome Evolution & Function, 25 Willcocks St., University of Toronto, Toronto, Ontario, Canada.
*These authors contributed equally to this work.

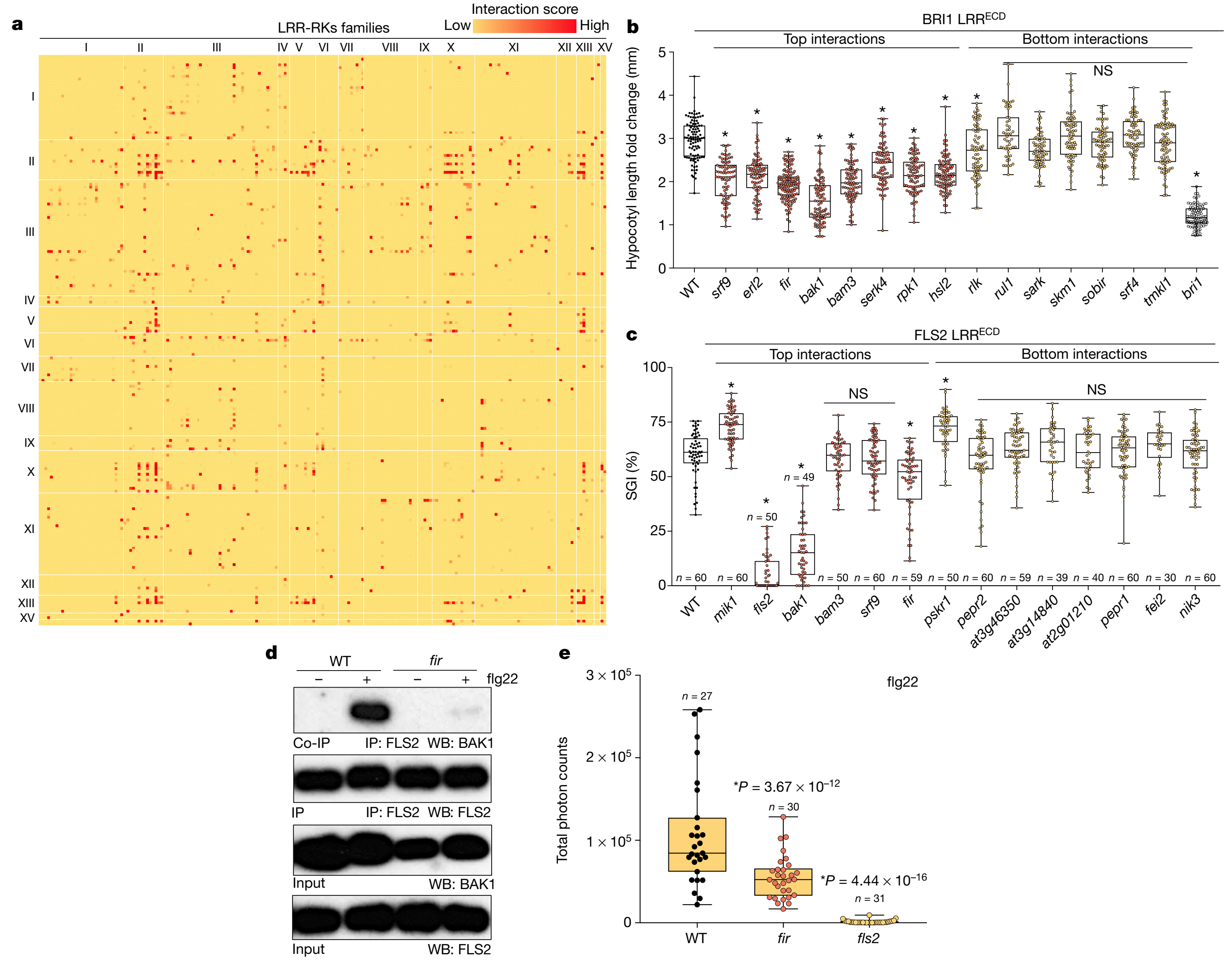
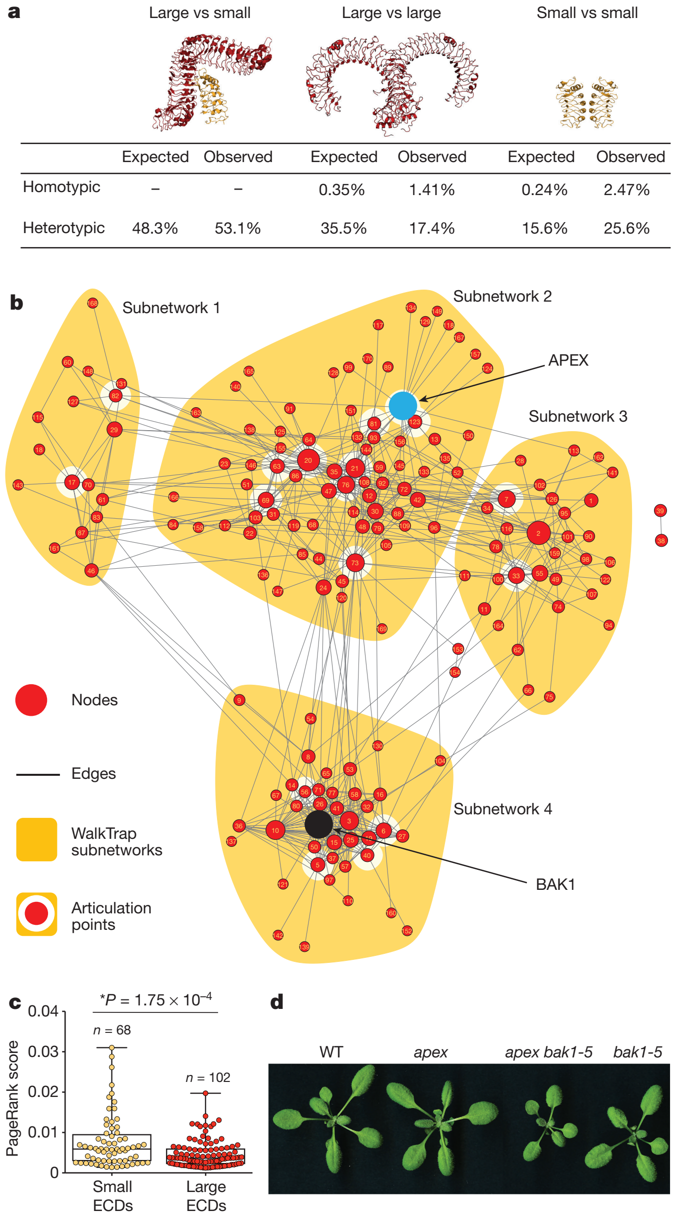


Figure 1 | CSI$^{\mathrm{LRR}}$ interaction map and functional validation. **a**, Interaction heat map organized by phylogenetic subgroups of LRR-RKs (roman numeral, XIV and XV are merged)[5]. The colour scale bar shows interaction score values. **b**, Hypocotyl length ratios of seedlings grown in the presence or absence of 500 nM brassinolide. See Methods for genotypes and Supplementary Methods for *n*, the number of biologically independent hypocotyls for all genotypes. WT, wild type. $*P = 3.17 \times 10^{-3}$ (*rlk* compared to wild type), $*P = 3.2 \times 10^{-15}$ (all others compared to wild type). NS, not significant. **c**, flg22-induced seedling growth inhibition (SGI). *n* denotes numbers of biologically independent seedlings. $*P = 3.14 \times 10^{-12}$ (*mik1*); $*P = 2.8 \times 10^{-15}$ (*fls2*), $*P = 2.8 \times 10^{-15}$ (*bak1*), $*P = 2.88 \times 10^{-10}$ (*fir*), $*P = 2.88 \times 10^{-10}$ (*pskr1*). In **b** and **c**, wild type (black) and mutant lines targeting the HCI (top interactions; red) and LCI (bottom interactions; yellow) partners for BRI1 and FLS2 are indicated and ordered by decreasing interaction score from left to right. Dots denote individual observations from six independent experiments. Box plots display the first and third quartiles, split by the median; whiskers extend to include the maximum and minimum values. Statistical significance was determined using linear mixed effect modelling, and *P* values are from a post hoc unpaired two-sided *t*-test corrected with the Holm method for multiple testing. **d**, Western blot analyses of FLS2–BAK1 co-immunoprecipitations (co-IP) in seedlings treated with either water (−) or flg22 (+) for 10 min. Anti-BAK1 or anti-FLS2 antibodies were used to analyse lysates from the genotypes indicated. Experiment was repeated three times with similar results. **e**, flg22-induced oxidative bursts represented as total photon counts over 40 min. Genetic backgrounds are indicated. Dots represent individual observations from three independent experiments. Box plots and statistical significance are as in **b** and **c**. *n* denotes numbers of biologically independent leaf discs.

altered signalling outputs (8 out of 8 for BRI1; 3 out of 5 for FLS2), mutants for the LCI$^{\mathrm{BRI1/FLS2}}$ partners were mostly indistinguishable from wild-type plants (6 out of 7 for BRI1; 7 out of 8 for FLS2) (Fig. 1b, c, Extended Data Figs 4b, 5b–e). Thus, we successfully used CSI$^{\mathrm{LRR}}$ to identify functionally relevant interactions for BRI1 and FLS2, and have expanded the repertoire of LRR-RKs known to contribute to plant steroid signalling and flagellin-based immunity.

The LRR-RK AT2G27060 (hereafter named FLS2-INTERACTING RECEPTOR, FIR) also interacted with the FLS2 co-receptor BAK1 in CSI$^{\mathrm{LRR}}$, suggesting that FIR may influence the FLS2–BAK1 signalling complex *in vivo*. FLS2–BAK1 complex formation was reduced upon flg22 treatment in the *fir* mutant (Fig. 1d), and this correlated with a reduction in flg22-induced ROS burst and *FLG22-INDUCED RECEPTOR KINASE 1* (*FRK1*) gene expression (Fig. 1e, Extended Data Fig. 6a). We also measured flg22-induced root growth inhibition as well as resistance against the bacterium *Pseudomonas syringae* pv. *tomato* DC3000 (*Pto* DC3000), and found that both were significantly reduced in *fir* mutants (Extended Data Fig. 6b–d). Thus, FIR regulates FLS2 signalling and facilitates flg22-induced BAK1–FLS2 complex formation.

Next, we defined the key principles that govern interactions in CSI$^{\mathrm{LRR}}$. LRR-RKs have large (greater than 12 LRR repeats) or small (less than 12 LRR repeats) ECDs, and the sizes are typically associated



	Large vs small		Large vs large		Small vs small	
	Expected	Observed	Expected	Observed	Expected	Observed
Homotypic	–	–	0.35%	1.41%	0.24%	2.47%
Heterotypic	48.3%	53.1%	35.5%	17.4%	15.6%	25.6%

Figure 2 | CSI^{LRR} is defined by four distinct subnetworks and two critical nodes. a, Expected and observed percentages of interactions organized by interaction types and ECD sizes. The expected percentages were calculated assuming random interaction between observed proteins. **b**, WalkTrap subnetworks are shown in orange. The diameter of the nodes (red circles) is proportional to their PageRank score. Numbers in each node are detailed in Extended Data Fig. 7. Edges (black lines) show interactions between nodes. BAK1 and APEX are marked in black and cyan, respectively. Articulation points are surrounded by a white halo. **c**, Small ECDs have higher PageRank scores than large ECDs. *n* denotes numbers of independent nodes (dots). Statistical significance was determined by an unpaired two-sided *t*-test. Box plots contain the first and third quartiles, split by the median; whiskers extend to include the maximum and minimum values. **d**, Representative rosettes of 20 biologically independent 3-week-old *Arabidopsis* plants. Genetic backgrounds are indicated.

with roles in ligand perception or regulation, respectively[1,21]. We compared the experimental pattern of interactions between these groups to the expected distribution of interactions assuming random binding (Fig. 2a). The distributions between the subgroups significantly differed from each other ($P < 0.0001$, chi-square test), indicating that binding events between ECDs in CSI^{LRR} are not random. We observed a four- and tenfold overabundance of homotypic interactions between large and small ECDs, respectively (Fig. 2a), and also detected an increase in heterotypic interactions between small and large ECDs (Fig. 2a). We propose that plants have evolved small LRR-RKs to connect their otherwise unconnected larger counterparts.

Next, we used the WalkTrap algorithm and identified four LRR-RK subnetworks[22] (Fig. 2b, Extended Data Fig. 7, Supplementary Table 8), of which at least one is biologically relevant (Supplementary Text 4). The PageRank algorithm was then used to compare the contributions of small and large ECDs to CSI^{LRR} connectivity[23] (Supplementary Table 9). Nodes corresponding to small ECDs have significantly higher PageRank values and are thus more essential to the overall connectivity of the network (Fig. 2c). Notably, BAK1 (a small LRR-RK) was measured by PageRank as the most interconnected and important node in CSI^{LRR}.

Articulation points are nodes whose removal from a network results in the formation of at least two disconnected subnetworks[24]. Removal of the small LRR-RK AT5G63710 (hereafter named APEX) resulted in the loss of the most nodes from the core structure of CSI^{LRR}, and was thus defined as the most important articulation point for network integrity (Supplementary Table 10). We predicted that genetic elimination of *APEX* and *BAK1* would have obvious developmental consequences. To test this, we constructed *apex bak1-5* double-mutant plants[25]. Although *apex* and *bak1-5* single-mutant plants were morphologically wild type, *apex bak1-5* double-mutant plants were developmentally impaired (Fig. 2d). Thus, network properties defined *in silico* are relevant in living plants.

In our screen, APEX interacted with PEPR1 and PEPR2. To test whether APEX associates with PEPR1 or PEPR2 in the context of the full-length receptors, we performed co-immunoprecipitation assays. PEPR1 and PEPR2 both associated with APEX in plant cells in the presence or absence of the Pep2 peptide ligand[26] (Fig. 3a, b, Supplementary Fig. 1). We next investigated whether the gene dosage of *APEX* would alter PEPR1 or PEPR2 signalling (Extended Data Fig. 8a, b). *apex* knockout plants and two independent overexpression lines (*35S::APEX*) all displayed reduced Pep2-induced bursts of ROS (Fig. 3c). The further reduction in Pep2-triggered ROS bursts in *apex bak1-5* double-mutant plants indicates that both BAK1 and APEX are required for wild-type PEPR1 and PEPR2 signalling (Fig. 3c). Thus, APEX interacts with PEPR1 and PEPR2 in a ligand-independent manner, and a wild-type *APEX* dosage is required for appropriate Pep2-induced responses.

Next, we predicted that changes in *APEX* dosage would affect the function of CSI^{LRR} as a coherent structural unit *in vivo*, thereby affecting the function of receptors without a direct physical interaction. To test this concept, we analysed whether the functions of BRI1 and FLS2, two receptors that reside several network-steps away from APEX, were affected in our set of *APEX* lines. The overexpression of *APEX* had either inconsistent effects or no effects on brassinosteroid-induced hypocotyl elongation and flg22-induced ROS bursts (Fig. 4a, b, Extended Data Fig. 8c). By contrast, BRI1 and FLS2 functions were both altered in *apex* mutants, as indicated either by the low levels of hypocotyl elongation in response to brassinosteroid or by the enhanced flg22-induced ROS bursts (Fig. 4a, b, Extended Data Fig. 8d, e). Notably, these aberrant ligand-induced signalling responses were both dependent on BAK1 (Fig. 4a, b). Finally, *apex* mutants showed a notable increase in flg22-induced FLS2–BAK1 complex formation, mitogen-activated protein kinase (MAPK) activation and *FRK1* expression (Fig. 4c–e, Supplementary Fig. 1). Thus, elimination of *APEX* has destabilizing effects in otherwise well-balanced LRR-RK signalling pathways.

To support our contention that the removal of articulation points results in network disruption *in vivo*, we established that mutations in *AT5G51560* (another predicted articulation point in CSI^{LRR}) altered BRI1 function (Extended Data Fig. 9). Our analysis has defined

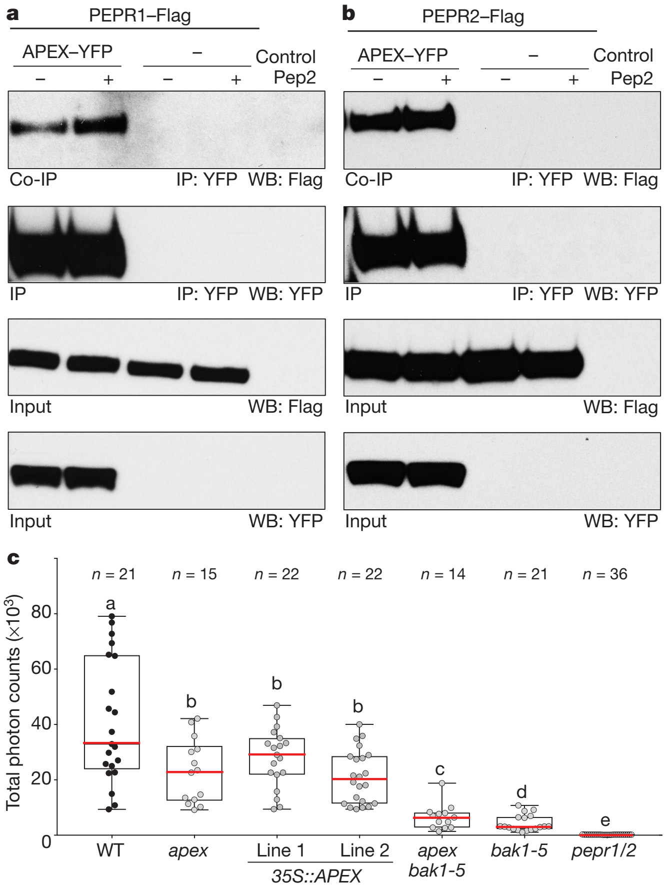


Figure 3 | APEX interacts with PEPR1 and PEPR2 to regulate danger peptide signalling. a, **b**, *Nicotiana benthamiana* leaves expressing Flag-tagged variants of PEPR1 or PERP2 either alone or together with a yellow fluorescent protein (YFP)-tagged APEX were treated with water (−) or Pep2 (+). Western blot analyses of PEPR1–APEX (**a**) and PEPR2–APEX (**b**) (co-)immunoprecipitations. Anti-Flag and anti-YFP antibodies were used to analyse lysates. These experiments were repeated three times with similar results. Full scans of the blots are in Supplementary Fig. 1. **c**, Pep2-induced oxidative bursts represented as total photon counts over 40 min. Genetic backgrounds are indicated. Dots represent individual observations from three independent experiments. *n* denotes numbers of biologically independent leaf discs. Box plots display the first and third quartiles, split by the median (red line); whiskers extend to include the maximum and minimum values. Statistical significance was determined by linear mixed effect modelling. The letters on top of the boxes (a–e) indicate the results of a post hoc Tukey test; genotypes with the same letter are indistinguishable at >95% confidence.

16 additional LRR-RKs as articulation points of CSI^{LRR} (Supplementary Table 10). Although the removal of any one of these leads to the fragmentation of CSI^{LRR} into no more than three subnetworks, these articulation points make tempting targets to study the LRR-RK family of receptors at the system level.

The mineable resources introduced here have provided insights into the wiring diagram that underlays LRR-RK signalling. We propose that LRR-RKs operate in a unified regulatory network governed by the following key guiding tenets: (i) ligand-induced signalling is modulated locally by the presence and/or activities of other LRR-RKs; (ii) small LRR-RKs, in addition to their function as co-receptors, act as regulatory scaffolds and organize their larger counterparts into a signalling network; and (iii) coupling of LRR-RK signalling to the overall stability of the network ensures appropriate response modulation by network-feedback mechanisms, an overlooked determinant of response specificity.

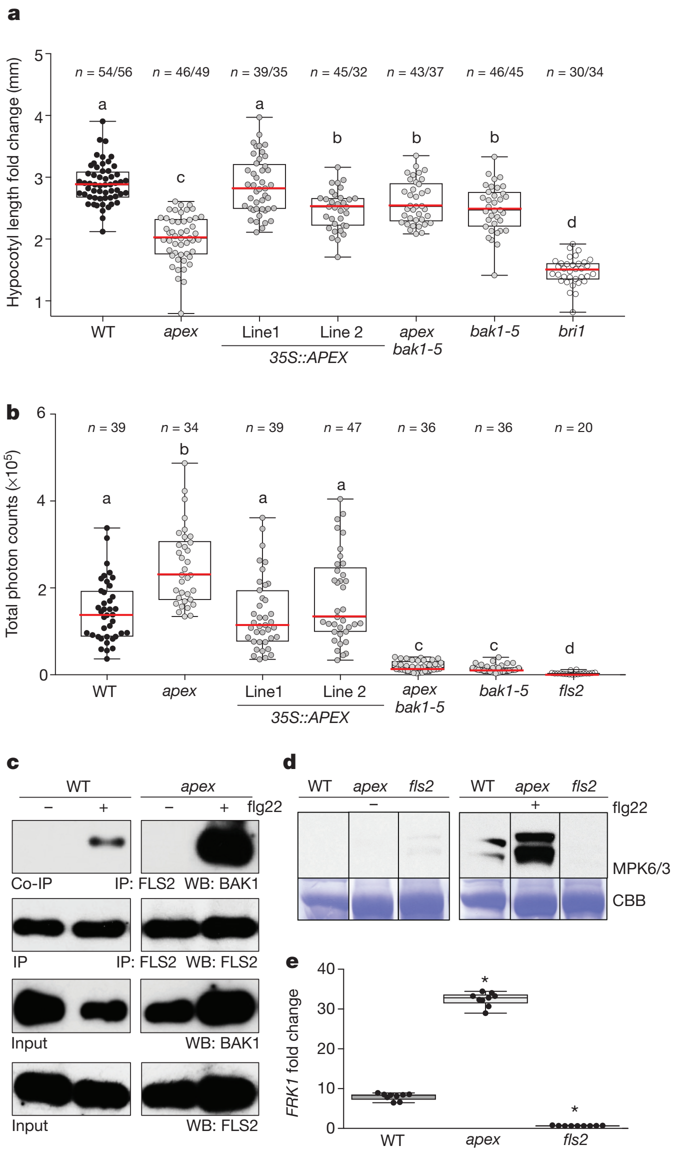


Figure 4 | CSI^{LRR} functions as a unified regulatory network. a, Hypocotyl length ratios of seedlings grown in the presence or absence of 500 nM brassinolide. Genotypes are indicated (wild type, black). Dots represent individual observations from three independent experiments. *n* denotes numbers of biologically independent hypocotyls in the presence/absence of brassinolide. **b**, flg22-induced oxidative burst in leaf discs of the genetic backgrounds indicated. Dots represent individual total photon counts over a 40-min time course; observations are from five independent experiments. *n* denotes the numbers of biologically independent leaf discs. Box plots in **a** and **b** display the first and third quartiles, split by the median (red line); whiskers extend to include the maximum and minimum values. Statistical significance was determined by linear mixed effect modelling; letters above the boxes (a–c) indicate the results of a post hoc Tukey test. Genotypes with the same letter are indistinguishable at >95% confidence. **c**, Western blot analyses of FLS2–BAK1 co-immunoprecipitations in seedlings treated with either water (−) or flg22 (+). An anti-BAK1 or anti-FLS2 antibody was used to analyse lysates. Experiment was repeated three times with similar results. **d**, flg22-induced activation of MAPKs in the genotypes indicated. The phosphorylated MPK3/6 proteins were detected with an anti-pERK antibody. Experiment was repeated four times with similar results. Colloidal brilliant blue (CBB) staining shows equal loading of the samples. Full scans of the blots in **c** and **d** are presented in Supplementary Fig. 1. **e**, Seedlings of the genotypes indicated were treated with either water or flg22 and changes in *FRK1* transcript levels were quantified by qPCR. Dots represent individual observations from three independent experiments. $n = 9$ biologically independent mRNA samples for all genotypes. Box plots are as in **a** and **b**. *$P < 0.05$, unpaired two-sided *t*-test followed by multiple testing correction using the Holm method.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 October 2016; accepted 29 November 2017.
Published online 10 January 2018.

1. Belkhadir, Y., Yang, L., Hetzel, J., Dangl, J. L. & Chory, J. The growth–defense pivot: crisis management in plants mediated by LRR-RK surface receptors. *Trends Biochem. Sci.* **39,** 447–456 (2014).
2. Jaillais, Y., Belkhadir, Y., Balsemão-Pires, E., Dangl, J. L. & Chory, J. Extracellular leucine-rich repeats as a platform for receptor/coreceptor complex formation. *Proc. Natl Acad. Sci. USA* **108,** 8503–8507 (2011).
3. Özkan, E. *et al.* An extracellular interactome of immunoglobulin and LRR proteins reveals receptor–ligand networks. *Cell* **154,** 228–239 (2013).
4. Braun, P. *et al.* An experimentally derived confidence score for binary protein–protein interactions. *Nat. Methods* **6,** 91–97 (2009).
5. Sun, J., Li, L., Wang, P., Zhang, S. & Wu, J. Genome-wide characterization, evolution, and expression analysis of the leucine-rich repeat receptor-like protein kinase (LRR-RLK) gene family in Rosaceae genomes. *BMC Genomics* **18,** 763 (2017).
6. Shiu, S. H. *et al.* Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. *Plant Cell* **16,** 1220–1234 (2004).
7. Hohmann, U., Lau, K. & Hothorn, M. The structural basis of ligand perception and signal activation by receptor kinases. *Annu. Rev. Plant Biol.* **68,** 109–137 (2017).
8. Zipfel, C. & Oldroyd, G. E. Plant signalling in symbiosis and immunity. *Nature* **543,** 328–336 (2017).
9. Soyars, C. L., James, S. R. & Nimchuk, Z. L. Ready, aim, shoot: stem cell regulation of the shoot apical meristem. *Curr. Opin. Plant Biol.* **29,** 163–168 (2016).
10. Song, W., Han, Z., Wang, J., Lin, G. & Chai, J. Structural insights into ligand recognition and activation of plant receptor kinases. *Curr. Opin. Struct. Biol.* **43,** 18–27 (2017).
11. Boutrot, F. & Zipfel, C. Function, discovery, and exploitation of plant pattern recognition receptors for broad-spectrum disease resistance. *Annu. Rev. Phytopathol.* **55,** 257–286 (2017).
12. Santiago, J., Henzler, C. & Hothorn, M. Molecular mechanism for plant steroid receptor activation by somatic embryogenesis co-receptor kinases. *Science* **341,** 889–892 (2013).
13. Sun, Y. *et al.* Structural basis for flg22-induced activation of the *Arabidopsis* FLS2–BAK1 immune complex. *Science* **342,** 624–628 (2013).
14. Sun, Y. *et al.* Structure reveals that BAK1 as a co-receptor recognizes the BRI1-bound brassinolide. *Cell Res.* **23,** 1326–1329 (2013).
15. Tang, J. *et al.* Structural basis for recognition of an endogenous peptide by the plant receptor kinase PEPR1. *Cell Res.* **25,** 110–120 (2015).
16. Braun, P. Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays. *Proteomics* **12,** 1499–1518 (2012).
17. Mukhtar, M. S. *et al.* Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science* **333,** 596–601 (2011).
18. Alonso, J. M. *et al.* Genome-wide insertional mutagenesis of *Arabidopsis. thaliana. Science* **301,** 653–657 (2003).
19. Belkhadir, Y. *et al.* Brassinosteroids modulate the efficiency of plant immune responses to microbe-associated molecular patterns. *Proc. Natl Acad. Sci. USA* **109,** 297–302 (2012).
20. Mott, G. A. *et al.* Genomic screens identify a new phytobacterial microbe-associated molecular pattern and the cognate *Arabidopsis* receptor-like kinase that mediates its immune elicitation. *Genome Biol.* **17,** 98 (2016).
21. Shinohara, H., Mori, A., Yasue, N., Sumida, K. & Matsubayashi, Y. Identification of three LRR-RKs involved in perception of root meristem growth factor in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **113,** 3897–3902 (2016).
22. Liu, W., Pellegrini, M. & Wang, X. Detecting communities based on network topology. *Sci. Rep.* **4,** 5739 (2014).
23. Li, X. Q., Xing, T. & Du, D. Identification of top-ranked proteins within a directional protein interaction network using the PageRank algorithm: applications in humans and plants. *Curr. Issues Mol. Biol.* **20,** 13–28 (2016).
24. Tian, L., Bashan, A., Shi, D. N. & Liu, Y. Y. Articulation points in complex networks. **8,** 14223, doi:10.1038/ncomms14223 (2017).
25. Schwessinger, B. *et al.* Phosphorylation-dependent differential regulation of plant growth, cell death, and innate immunity by the regulatory receptor-like kinase BAK1. *PLoS Genet.* **7,** e1002046 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Pasha for uploading the CSI^{LRR} interaction dataset to the Bio-Analytic Resource for Plant Biology, Y. Dagdas for comments on the manuscript, C. K. Garcia for the pECIA2/14 vectors, E. Özkan for his protocols, Y. Saijo, K. Tori and M. Butenko for the *pepr1/2*, *erl2* and *hsl2* mutant lines, respectively, and A. Bindeus for help with programming software. This work was supported by grants from the Austrian Academy of Sciences through the Gregor Mendel Institute (Y.B. and W.B.); the Natural Sciences and Engineering Research Council of Canada Discovery Grants to D.S.G. and D.D.; a Canada Research Chair in Plant-Microbe Systems Biology (D.D.) or Comparative Genomics (D.S.G.); and the Centre for the Analysis of Genome Evolution and Function (D.D. and D.S.G.). This research was also funded by the Gatsby Charitable Foundation (C.Z.) and the European Research Council (grant 'PHOSPHinnATE') (C.Z.). E.S.-L. is supported by a Hertha Firnberg Programme post-doctoral fellowship (T-947) from the FWF Austrian Science Fund. M.S. is supported by a post-doctoral fellowship (STE 2448/1) from the Deutsche Forschungsgemeinschaft (DFG). This work was supported by the National Science Foundation (IOS-1557796) to M.S.M. P.S.-B. acknowledges funding from the Austrian Federal Ministry of Science, Research & Economy, and the City of Vienna through the Vienna Biocenter Core Facilities (VBCF). We would like to thank P. Serrano Drozdowskyj, A. Aszodi and A. Gyenesei from the VBCF Biocomputing facility for developing the Platero software. We also thank the VBCF Plant Sciences facilities for the plant growth chambers.

Author Contributions E.S.-L., G.A.M., K.G. and Y.B conceived and designed the experiments for the CSI screen. J.N. and A.L. cloned and expressed all the ECDs with inputs from E.S.-L., P.S.-B. and Y.B.; E.S.-L. performed all the ECD interaction assays. T.C.H. conceived, designed and performed the Y2H assays under the supervision of M.S.M.; G.A.M., E.S.-L. and K.P. characterized and tested all the T-DNA insertion lines under the supervision of D.D., D.S.G. and Y.B.; G.A.M., E.S.-L., K.P., N.W., K.G. and J.K. genotyped and bulked all the T-DNA insertion lines. G.A.M. analysed and implemented the computational and statistical analysis of all the data with inputs from D.S.G. and Y.B.; M.L. and G.A.M. conceived, designed and performed the network analysis with inputs from D.D., D.S.G., N.J.P. and Y.B.; M.S. designed and performed the BAK1–FLS2 co-immunoprecipitation assays under the supervision of C.Z.; K.P. organized and performed the APEX–PEPRs co-immunoprecipitation experiment with guidance from E.S.-L.; E.S.-L. and K.G. generated the *apex bak1* double-mutant and *35S::APEX* transgenic lines. S.B.S. and W.B. contributed and characterized the *fir* T-DNA mutant. E.S.-L. conceived, organized and performed the physiological assays with brassinosteroids, Pep2 and flg22. D.M. and M.M. devised the synthesis of the flg22 and Pep2 peptides. G.A.M. and Y.B. wrote the manuscript with major input from E.S.-L., M.S.M., C.Z., D.D. and D.S.G.; all authors commented and agreed on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to Y.B. (youssef.belkhadir@gmi.oeaw.ac.at), D.S.G. (david.guttman@utoronto.ca) or D.D. (darrell.desveaux@utoronto.ca).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Molecular cloning of LRR-RK extracellular domains. For each ECD cloned, we determined the boundaries of signal peptides and transmembrane domains using a range of bioinformatics tools[27]. A key step in defining the boundaries of each ECD was the identification of the N- and C-terminal cysteine-capping consensus motifs (CXXXXC and variations thereof) that border most of the *Arabidopsis* ECDs. This was achieved by visual inspection of the primary amino acid sequences. These cysteine caps are thought to cap the exposed edges of the hydrophobic core formed by the repetition of the LRRs and produce disulfide bonds that preserve the tertiary protein structure. We found that the cysteine caps were important for enhancing ECD solubility and preventing aggregation and proteolysis *in vitro*. For expression in *Drosophila melanogaster* Schneider 2 (S2) cells, each ECD was inserted into the pECIA-2 and the pECIA-14 expression vectors (a gift from C. K. Garcia)[3]. pECIA2/14 are derivatives of the pMT/BiP/V5 (Invitrogen, V4130-20), which uses a copper-inducible *Drosophila* metallothionein promoter and have the signal sequence of the *Drosophila* BiP protein. The ECDs were cloned by sequence and ligation independent cloning (SLIC) between the existing BiP signal sequence and the C-terminal epitope tags specific to each vector. Sanger sequencing confirmed the presence of each insert. Primers were designed to have a sequence partially homologous to the desired boundaries of the ECDs followed by extensions for RecA-mediated SLIC strategy attached (Supplementary Table 1). Amplification was done using Phusion Flash Mastermix Thermo Fisher Scientific according to the manufacturer's instructions for 2-step PCR. ECDs (176 out of 200) were cloned from plasmid templates available from the *Arabidopsis* Biological Resource Center (ABRC)[28]. Twenty-four ECDs were cloned from *Arabidopsis* seedlings and mature leaves using RT–PCR, followed by amplification as described above and by RecA-mediated SLIC.

Secreted expression of LRR-RK extracellular domains. The ECDs cloned into the pECIA2 (for expression as a bait) and pECIA14 (for expression as a prey) vectors were expressed using transient transfection of *Drosophila* S2 cells cultured at 27 °C. Upon transfection using Effectene (Qiagen), the culturing temperature was changed to 21 °C. Twenty-four hours after transfection, protein expression was induced with 1 mM $CuSO_4$ and supernatant was collected three days after induction. Protease inhibitors (Sigma) and 0.02% NaN_3 were added to the medium (ESF 921, Expression Systems) containing the recombinant ECDs and then stored at 4 °C before use. The cell supernatant was assessed for recombinant protein expression by western blotting using anti-V5 antibodies (Invitrogen) for the baits or by alkaline phosphatase activity quantification for the preys.

CSI^{LRR} primary screen. Pairwise interaction assays were performed as detailed previously[3] for the extracellular interactome assay (ECIA) with the slight modifications indicated below. Schneider's medium containing recombinant ECDs was subjected to a fourfold dilution in a PBS buffer containing 1 mM CaC_{l2}, 1 mM $MgCl_2$ (equilibration buffer) and 0.1% bovine serum albumin (BSA; Sigma). Bait proteins fused to the Fc domain were captured directly on 96-well protein-A-coated plates (Thermo Fisher Scientific) by overnight incubation at 4 °C. Protein-A-coated plates were washed in a PBS solution containing 0.1% Tween-20 before use. The bait-coated plates were blocked with the equilibration buffer containing 1% BSA for 3 h at 4 °C and subsequently washed. The prey proteins fused to the alkaline phosphatase were then added to the wells and incubated for 2 h at 4 °C and then washed away before adding the alkaline phosphatase substrate (KPL 50-88-02). Upon addition of the substrate, plates were incubated for 2 h at room temperature and alkaline phosphatase activity was monitored by measuring the absorbance at 650 nm using a Synergy H4 Multi-Mode plate reader (BioTek). Images of the 96-well plates were acquired for visual inspection. The complete set of raw absorbance values was combined into a binary dataset using an in-house-designed script (Platero v0.1.4), and then subjected to post experimental statistical analysis to remove both false positive and false negative interactions.

CSI^{LRR} data analysis. The complete set of absorbance values for each pairwise interaction was combined into a data matrix. To make measurements comparable across plates and eliminate any bias in the data arising from the differential background binding capacities of the baits and preys we used a two-way median polish[29,30]. The residuals were then used to calculate the median and median absolute deviation (MAD). The MAD is the median of the absolute values of the residuals (deviations) from the data's median. The MAD was used for the calculation of modified *Z*-scores for each individual interaction measured. The modified *Z*-score used here is (i) nonparametric and makes minimal distributional assumptions, (ii) minimizes measurement bias due to positional effects and (iii) is resistant to statistical outliers. The modified *Z*-score usually excludes control measurements altogether under the assumption that most interactions in a screen such as CSI^{LRR} would be unproductive and thus serve as controls. However, during the primary screen each 96-well plate contained two mock prey negative control wells and one well with the positive control interaction pair BAK1–BIR4[31]. To identify high-stringency bidirectional interactions, we calculated the geometric mean modified *Z*-score of the interaction as measured in the bait–prey and prey–bait orientations. Any value for which the geometric mean product of the *Z*-scores was greater than 2.5 was considered significant for the purposes of network construction.

CSI^{LRR} retest screens. All of the HCI in CSI^{LRR} and a randomly selected subset of LCI^{CSI} were independently retested. Each ECD was newly expressed and all retested interactions were assayed in both bait–prey orientations. For each interaction tested, three prey-only negative control wells were included, to control for non-specific binding. Thus, a total of six negative controls were tested for each bidirectional interaction. One well containing the positive control interaction pair BAK1–BIR4 was included on each plate. The two-way median polish and modified *Z*-scoring system used in the initial screen depends upon large numbers of non-interactions to perform reliably. The low sample number, enriched with high or low performing protein pairs, led to an asymmetrical data distribution in the retest, making it inappropriate to implement our original hit calling method. Instead, we implemented a multi-stage hit calling process to ensure reliable data confirmation. The absorbance values were paired with the corresponding value from the CSI^{LRR} and subjected to an interquartile range normalization step to ensure the two datasets could be accurately compared (Extended Data Fig. 3). The geometric mean of the normalized absorbance values for each bidirectional interaction was then calculated. The threshold for inclusion in the positive interaction set was set to the lowest geometric mean absorbance value found in the 567 interactions present in the CSI^{LRR} (absorbance value = 0.090989). Therefore, any interaction with a geometric mean absorbance value > 0.090989 was considered positive; all others were considered negative.

CSI^{LRR} network construction and analysis. The network was constructed using the igraph package (http://igraph.org/r/) in the R programming environment (https://www.r-project.org/). To identify clusters of interacting proteins in the network, we used the WalkTrap algorithm as implemented in igraph; this algorithm is based on the concept that if one performs random walks on a network, then the walks are more likely to stay within the densely connected parts of the network, thus corresponding to clusters with higher levels of interconnectedness[22]. The WalkTrap was implemented with edge weights corresponding to the interaction score and a length of random walk of 8. To measure the importance of each node within the network, we applied the PageRank algorithm as implemented in igraph, which operates by counting the number and quality of links to a node, thereby establishing its importance and assigning a 'weight' value to it[23]. In simpler terms, PageRank measures node connectivity via the number of connections to other nodes. The PageRank algorithm is an example of a link-analysis algorithm, which are iterative and interactive data analysis techniques that operate with the underlying assumption that nodes with higher scores are likely to be more connected to other nodes when compared to nodes with lower scores[23]. The PageRank implementation using the PRPACK library within the igraph package was used with edge weights corresponding to the interaction score, and a damping factor of 0.85, which is also the default. Finally, we identified the articulation points (or cut vertices) in the network. An articulation point is any node in a unidirectional network the removal of which disconnects the network.

Y2H assays with LRR-RK ICDs. The Y2H experiment was conducted as described previously[17] with some modifications. In brief, we used a collection of LRR-RK ICDs cloned in both bait and prey plasmids[17]. The ICDs of the LRR-RKs were fused to the GAL4 activation domain using a pDEST-AD-CHY2 vector with a tryptophan selection marker to form the prey constructs and to the GAL4 DNA-binding domain using a pDEST-DB vector with a leucine selection marker to form the bait constructs. Target prey and bait constructs were transformed into *Saccharomyces cerevisiae* strains Y8800 (MATa) and Y8930 (MATα), respectively. Transformations were confirmed by selecting the haploid yeast strains on their corresponding selective medium (SD-T and SD-L). Haploid bait and prey strains were mated in liquid YEPD (yeast extract 10 g l^{-1}, peptone 20 g l^{-1}, dextrose 20 g l^{-1}, adenine 100 mg l^{-1}) medium overnight at 30 °C. The resulting diploid yeasts were selected in liquid SD-LT medium for 48 h at 30 °C. Reconstitution of the GAL4 transcription factor through the interaction of the bait and prey led to the activation of a HIS3 reporter gene and subsequently biosynthesis of histidine. Because the pDEST-AD vector contains the *CHY2* (a cycloheximide-sensitive gene), any growth on the yeast medium containing cycloheximide constitutes a false positive interaction. Equal amounts of diploid yeasts were transferred to solid SD-LTH (positive selection plates) and SD-LH+ cycloheximide (20 mg l^{-1}) medium (*de novo* auto-activation plates). Interactions were scored positive if there was growth on positive selection plates, but no growth on *de novo* auto-activation plates. The retest on the random LCI^{Y2H} pairs was performed in similar conditions.

T-DNA insertions of top and bottom BRI1- and FLS2-interaction partners. Noting that our statistical cut-off for considering an interaction for network construction was set to a CSI score (geometric mean modified Z-score) > 2.5, we compiled a list of 'top-interactions' (HCI$^{BRI1/FLS2}$; CSI score > 1.75) and 'bottom-interactions' (LCI$^{BRI1/FLS2}$; CSI score = 0) (Supplementary Table 6). We amassed a collection of T-DNA insertion lines from the *Arabidopsis* Biological Resource Centre (ABRC) for the HCI$^{BRI1/FLS2}$ and LCI$^{BRI1/FLS2}$ genes, focusing when possible on exon insertions closest to the 5′ end of each gene. For each of the mutant lines we performed PCR tests for the presence of non-segregating (homozygous) T-DNA insertions in each target gene as well as qPCR analysis of altered target gene expression (Extended Data Figs 4, 5, Supplementary Table 7). For BRI1, we tested mutant lines targeting the following interaction partners: HCI^{BRI1} top genes; first rank: *STRUBBELIG-RECEPTOR FAMILY 9* (*SRF9*)[32], second rank: *ERECTA-LIKE 2* (*ERL2*)[33], third rank: *FIR/AT2G27060* (this study), fourth rank: *BAK1* (*bak1-4 allele*)[25,34,35], sixth rank: *BARELY ANY MERISTEM 3* (*BAM3*)[36], seventh rank: *SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASE 4* (*SERK4*)[35,37], eighth rank: *RECEPTOR-LIKE PROTEIN KINASE 1* (*RPK1*)[38] and ninth rank: *HAESA-LIKE 2* (*HSL2*)[39–41]. We were not able to test the following genes; fifth rank: *RECEPTOR-LIKE PROTEIN KINASE 2* (*RPK2*)[42] and tenth rank: *BAK1 INTERACTING RECEPTOR 4* (*BIR4*)[31]. The mutant lines obtained from the Salk Institute (La Jolla; the SALK lines collection, http://signal.salk.edu/cgi-bin/tdnaexpress) were annotated as homozygous for the T-DNA inserts but we genotyped both as wild-type plants. LCI^{BRI1} bottom genes; 191st rank: *RECEPTOR-LIKE KINASE* (*RLK*)[43], 193rd rank: *REDUCED IN LATERAL GROWTH1* (*RUL1*)[44], 194th rank: *SENESCENCE-ASSOCIATED RECEPTOR-LIKE KINASE* (*SARK*)[45], 196th rank: *STERILITY REGULATING KINASE MEMBER 1* (*SKM1*)[46], 197th rank: *SUPPRESSOR OF BIR1-1* (*SOBIR1*)[47], 198th rank: *STRUBBELIG-RECEPTOR FAMILY 4* (*SRF4*)[32] and 200th rank: *TRANSMEMBRANE KINASE LIKE 1* (*TMKL1*)[48]. The following genes were not tested; 192th rank: *RECEPTOR-LIKE KINASE 902* (*RLK902*)[49] and 199th rank: *TRANSMEMBRANE KINASE 1* (*TMK1*)[50]. Although annotated as homozygous for the T-DNA insert in the SALK database, we genotyped both lines as wild-type plants.

For FLS2, we tested mutant lines targeting the following HCI^{FLS2} top genes; first rank: *MDIS1-INTERACTING RECEPTOR LIKE KINASE 1* (*MIK1*)[51], second rank: *FLS2* as an internal control but also as a self-interaction[52], third rank: *FIR/AT2G27060* (this study), fifth rank*: BAK1* (*bak1-4 allele*)[25,34,35], seventh rank: AT5G62710, eighth rank: *BARELY ANY MERISTEM 3* (*BAM3*)[36], 13th rank: *RECEPTOR-LIKE PROTEIN KINASE 1* (*RPK1*)[38], 14th rank*: STRUBBELIG-RECEPTOR FAMILY 9* (*SRF9*)[32] and 15th rank: AT2G27060. We did not test the following; fourth rank*: ERECTA* and sixth rank*: ERECTA-LIKE 2* (*ERL2*) because the *er* mutant shows altered flg22-induced marker gene expression[33], and tenth rank: *IMPAIRED-OOMYCETE SUSCEPTIBILITY 1* (*IOS1*), which has been implicated in flg22-induced ROS burst, marker gene expression and FLS2–BAK1 complex formation[53]. The following T-DNA lines were not tested; fifth rank*: RECEPTOR-LIKE PROTEIN KINASE 2* (*RPK2*)[42], eleventh rank*: RECEPTOR-LIKE KINASE 1* (*RKL1*)[49] and twelfth rank*: BAK1 INTERACTING RECEPTOR 4* (*BIR4*)[31] because we genotyped them as wild type despite their annotation as homozygous for the presence of a T-DNA insert. LCI^{FLS2} bottom genes; 190th rank*: phytosulfokine peptide receptor 1* (*PSKR1*)[54–56], 191st rank*: PEPR2*[57], 192nd rank*: AT3G46350*, 194th rank: *AT3G14840*, 195th rank: *AT2G01210*, 196th rank: *PEPR1*[57], 198th rank: *FEI2*[31] and 200th rank*: NSP-INTERACTING KINASE 3* (*NIK3*)[58].

Brassinosteroid hypocotyl responses assays. These assays have been performed as described previously[2,19].

Peroxidase flg22 responses assays. The peroxidase assay was carried out as described previously[20]. In brief, leaf discs were taken from 4-week-old *A. thaliana* plants. The discs were washed for 1 h in 1 ml of 1× MS solution with agitation. After washing, discs were transferred to individual wells of a clear 96-well assay plate avoiding the use of the edge wells to minimize evaporation effects. Each well received 50 μl of 1× MS buffer alone, or supplemented with 1 μM of flg22 peptide. Plates were sealed with parafilm and incubated for 20 h with agitation. The leaf discs were removed and each well received 50 μl of a 1 mg ml^{-1} solution of 5-aminosalicylic acid (A79809, Sigma-Aldrich) pH 6.0 with 0.01% hydrogen peroxide. The reaction proceeded for 1–3 min and was stopped by the addition of 20 μl 2 N NaOH before reading the OD$_{600\ nm}$ on a POLARstar OPTIMA microplate reader (BMG Labtech). The flg22 peptide was obtained from Genscript.

Transient expression in *Nicotiana benthamiana*. *Agrobacterium tumefaciens* GV3101 strains were grown in LB medium supplemented with appropriate antibiotics overnight. Cultures were spun down and re-suspended in 10 mM $MgCl_2$ to OD$_{600\ nm}$ = 0.1. Agrobacterium strains carrying the pB35GWF binary plant expression vector for the expression of the full-length coding regions of PEPR1 (S1G73080BFF)[28] and PEPR2 (S1G17750BFF)[28] fused to a C-terminal Flag epitope tag were constructed and used for immunoprecipitation and western blot assays. pDONR-Zeo vector (Life Technologies) containing the cDNA of *APEX* (N5G63710ZEF) was used for gateway recombination in the binary plant expression vector pEarleyGate101 vector to generate the C-terminal YFP–HA tag fusion vector expressing *APEX-YFP-HA* under the control of the CaMV35S promoter. For each of the protein interaction pairs tested, the respective sets of agrobacterium strains were mixed 1:1 and syringe infiltrated into 3-week-old *N. benthamiana* leaves of plants grown in short-day conditions (12 h light:12 h dark). Samples for protein extraction were collected three days after infiltration before flash-freezing in liquid nitrogen.

Protein extraction and immunoprecipitation in *N. benthamiana*. Leaves were ground in liquid nitrogen and extraction buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 10% glycerol, 10 mM DTT, 10 mM EDTA, 1 mM NaF, 1 mM Na_2MoO_4•$2H_2O$, 1× (v/v) cOmplete Tablets, EDTA-free (Roche), and 1% (v/v) IGEPAL CA-630 (Sigma-Aldrich)) was added at 2 ml per gram tissue powder. Samples were homogenized by alternate rounds of Polytron and incubated in extraction buffer for 1 h at 4 °C. Samples were the clarified by a 20-min centrifugation step at 4 °C and 16,000*g*. Supernatants (3 ml) were adjusted to 2 mg ml^{-1} protein and incubated for 3 h at 4 °C with 30 μl GFP Trap-A beads (Chromotek) with slow but constant rotation. Following incubation, beads were washed four times with washing buffer containing 50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1% PMSF, and 0.1% IPEGAL. One hundred microlitres of 5× SDS Laemmli buffer was added to the beads, and the beads were heated at 95 °C for 10 min and subjected for further SDS–PAGE and immunoblotting analysis.

Plant cultivation, transgenic plants and mutants. The wild type used in all experiments was *A. thaliana* accession Columbia (Col-0). Unless specified otherwise, the *apex-1* allele was used in this work (Extended Data Fig. 8a). Plants were grown on soil or vertically on Petri dishes containing 0.5× Murashige and Skoog medium in long-day light conditions (16 h light:8 h dark). For *Pto* DC3000 pathogen assay and callose deposition upon flg22 treatment, plants were grown in short-day conditions (12 h light:12 h dark). The mutant plant genotypes used in this work are listed in Supplementary Table 7. For overexpression studies, the *35S::APEX-YFP-HA* construct was transformed separately into wild-type plants and more than 20 independent T$_1$ lines were isolated and between three and eight representative mono-insertion lines were selected in the T$_2$ generation. DNA genotyping, epifluorescence microscopy and protein extraction were performed on segregating T$_2$ to obtain homozygous T$_3$ generation lines with maximal expression levels (Extended Data Fig. 8b). The double-mutant *apex-1 bak1-5* was generated by crosses and genotyped for homozygosity using allele-specific primers for *apex-1* and dCAPS marker for *bak1-5* as described previously[25]. Genotyping was repeated for two consecutive generations and confirmed by Sanger sequencing. Primers are listed in Supplementary Table 7.

Protein extraction and immunoprecipitation in *Arabidopsis*. Fifteen to twenty seedlings were grown in each well of a 6-well plate for 2 weeks. Subsequently, seedlings were transferred to water and incubated overnight. The next day, flg22 was added at a final concentration of 100 nM and incubated for 10 min. Seedlings were than frozen in liquid nitrogen and subjected to protein isolation. To analyse FLS2–BAK1 receptor complex formation, proteins were isolated in 50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 10% glycerol, 5 mM dithiothreitol, 1% protease inhibitor cocktail (Sigma-Aldrich), 2 mM Na_2MoO_4, 2.5 mM NaF, 1.5 mM activated Na_3VO_4, 1 mM phenylmethanesulfonyl fluoride and 1% IGEPAL. For immunoprecipitations, anti-rabbit Trueblot agarose beads (eBioscience) coupled with anti-FLS2 antibodies and incubated with the crude extract for 2–3 h at 4 °C. Subsequently, beads were washed three times with wash buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM phenylmethanesulfonyl fluoride and 0.5% IGEPAL) before adding Laemmli sample buffer and incubating for 10 min at 95 °C. Analysis was carried out by SDS–PAGE and western blots using anti-FLS2 and anti-BAK1 antibodies[25].

Protein analysis. In all our protein manipulations, equal loading was ensured by Bradford protein quantification before loading and by CBB or Red Ponceau staining of the membrane post-protein transfer. Anti-GFP-HRP (MACS) and anti-Flag-HRP (Sigma-Aldrich) antibodies were used according to manufacturer's instructions. Polyclonal anti-FLS2 and anti-BAK1 antibodies were used as described previously[25]. Signal detection was achieved through chemiluminescence (SuperSignal West Pico Chemiluminescent Substrate, Thermo Fisher Scientific) and detected using autoradiography films (CL-XPosure Film, Thermo Fisher Scientific).

RNA isolation, cDNA synthesis and qPCR analysis. Total RNA was isolated from 1-week-old seedlings grown on 1/2 MS plates using either the GeneMATRIX Universal RNA Purification Kit (EURx) or TRI Reagent (Sigma-Aldrich), followed by DNaseI treatment (Thermo Fisher Scientific). Reverse transcription reactions were performed using up to 2 μg of total RNA and a reverse transcription kit (Applied Biosystems or Life Technologies). The cDNAs were used as a template for qPCR. qPCR analyses were performed using a Roche LightCycler96 instrument (Roche Applied Science) and data were analysed using the LightCycler 96 version

1.1 software or BioRad C1000 thermal cycler (BioRad). FastStart Essential DNA Green Master (Roche) or Maxima SYBR Green/ROX qPCR Master Mix (Thermo Fisher Scientific) were used according to manufacturer's instructions. Material from wild type plant served as the calibrator, and *ACTIN* or *UBQ10* was used as a reference. Relative gene expression levels were calculated using the $2^{-\Delta\Delta C_t}$ method. The amplification protocol consisted of: 95 °C for 1 min, (95 °C for 10 s, 55–62 °C for 10 s, 72 °C for 20 s) × 44 cycles. The relative mRNA levels were determined by normalizing the PCR threshold cycle number with *ACTIN* or *UBQ10*. All experiments were repeated three times independently, and the mean was calculated. The specificity of the amplification products was verified by melting curve analysis.

MAMP and DAMP responses assays. flg22 (QRLSTGSRINSAKDDAAGLQIA) and pep2 (DNKAKSKKRDKEKPSSGRPGQTNSVPNAAIQVYKED) peptides were synthesized at >95% purity by the in-house protein chemistry facility and dissolved to a 10 mM stock in pure water. For ROS burst assays, leaf disks (diameter 6 mm) were cut out from 4–5-week-old plants. Single disks were placed adaxial side up into 96-well microtitre plates in which every well contained 200 μl sterile MonoQ water. Floated disks were then vacuum infiltrated for 10 min. The plates were incubated on a rocking table at 45 r.p.m. in continuous light, at 21 °C for 5 h. The mix for elicitation containing 9.91 ml of MonoQ water, 40 μl of 500× HRP, 40 μl of 500× L-012 and appropriate peptide at a final concentration of 1 μM was freshly prepared in falcon tubes wrapped with aluminium foil on ice. For each well the water was carefully removed and replaced immediately with 100 μl of elicitation mix using a multichannel pipette. Relative luminescence measurements were started immediately after adding the elicitation mix using a BiTec Synergy 4 microplate reader. Horseradish peroxidase (HRP) was purchased from Sigma-Aldrich and prepared at a 10 mg ml^{-1} (500×) concentration in sterile MonoQ water. L-012 was purchased from Wako Chemicals GmbH. Preparation of a 500× L-012 stock solution containing 17 mg ml^{-1} L-012 in sterile MonoQ water and was subsequently protected from light. Solutions were stored at −20 °C. For the analysis of ROS burst data, the models were constructed using the total relative light units measured for the first 39 time points to ensure comparability across experiments. Root inhibition ratios were calculated on 7-day-old seedlings grown on plates left untreated or treated with 1 μM flg22.

Seedling growth inhibition assay. Seedlings of the noted *A. thaliana* lines were grown for 5 days on MS-agar plates with 1% sucrose before transfer of up to 10 seedlings to each well of a 6-well plate containing 1 ml of 0.5× MS medium with 1% sucrose. The seedlings were treated with water or 100 nM flg22 peptide and grown further for 7 days. The seedlings were removed, briefly dried, and weighed (fresh weight). The percentage of seedling growth inhibition was calculated by dividing the weight of individual treated seedlings by the mean weight of the control non-treated seedlings of the same genotype. The percentage of seedling growth inhibition was calculated by dividing the weight of individual treated seedlings by the mean weight of 10 non-treated seedlings of the same genotype. A maximum of 10 seedlings of each genotype were treated and the experiment was performed six times.

Pathogens assays. Assays with *Pseudomonas syringae* pv. *tomato* DC3000 (*Pto DC3000*) have been previously described[19]. Bacterial growth in plant leaves was assessed by inoculating 4-week-old plants with a bacterial inoculum of 10^5 colony-forming units (cfu) ml^{-1}. Growth inhibition of *Pto DC3000* by 1 μM flg22 was conducted as described[19]. Leaves were either infiltrated with water or with an elicitor solution containing 1 μM flg22. For each sample, four leaf discs were pooled and three samples were taken per data point (12 leaf discs in total). Leaf discs were bored from the infiltrated area and ground to homogeneity in 10 mM $MgCl_2$. The bacterial titre was determined by plating and serial dilution.

Program used for modelling and statistical analysis. Statistical analysis was performed using linear mixed effect modelling in the R programming environment (https://www.r-project.org/). Before modelling, data from independent experiments were combined and outliers were removed using the ROUT method, as implemented in GraphPad PRISM 7.0 (Q = 0.1%) (GraphPad Software, http://www.graphpad.com). Each dataset was checked for normality to ensure accurate modelling. qPCR data were analysed as fold induction, whereas all other data were $\log_{10}$ transformed before modelling to improve fit. Linear mixed effect models were constructed using the lme4 package: https://cran.r-project.org/package=lme4, using the genotype as a fixed effect and the individual experiment as a random effect. The resulting models were inspected for fit and further outlier checks were accomplished by examining both the Cook's distance and dfbeta distributions using the LMER Convenience Functions and influence.ME packages (https://CRAN.R-project.org/package=LMERConvenienceFunctions; https://cran.r-project.org/web/packages/influence.ME/). Statistical significance was determined using the lmerTest package (https://cran.r-project.org/package=lmerTest) using the Satterthwaite approximation and the resulting *P* values were corrected for multiple testing using the Holm method. In cases where pairwise comparisons were required, the adjusted *P* values were calculated using Tukey's honest significant difference (HSD) as implemented in the multcomp package (https://cran.r-project.org/package=multcomp).

To calculate the expected binding frequencies of a random network, we classified each node based on its ECD. Assuming equal frequency of a given node binding to any other node, the frequency for each class of binding event was calculated and divided into self-interactions between small ECDs (small–small homotypic), self-interactions between large ECDs (large–large homotypic), interactions between two different small ECDs (small heterotypic), interactions between two different large ECDs (large heterotypic), and interactions between one small and one large ECD (small–large heterotypic).

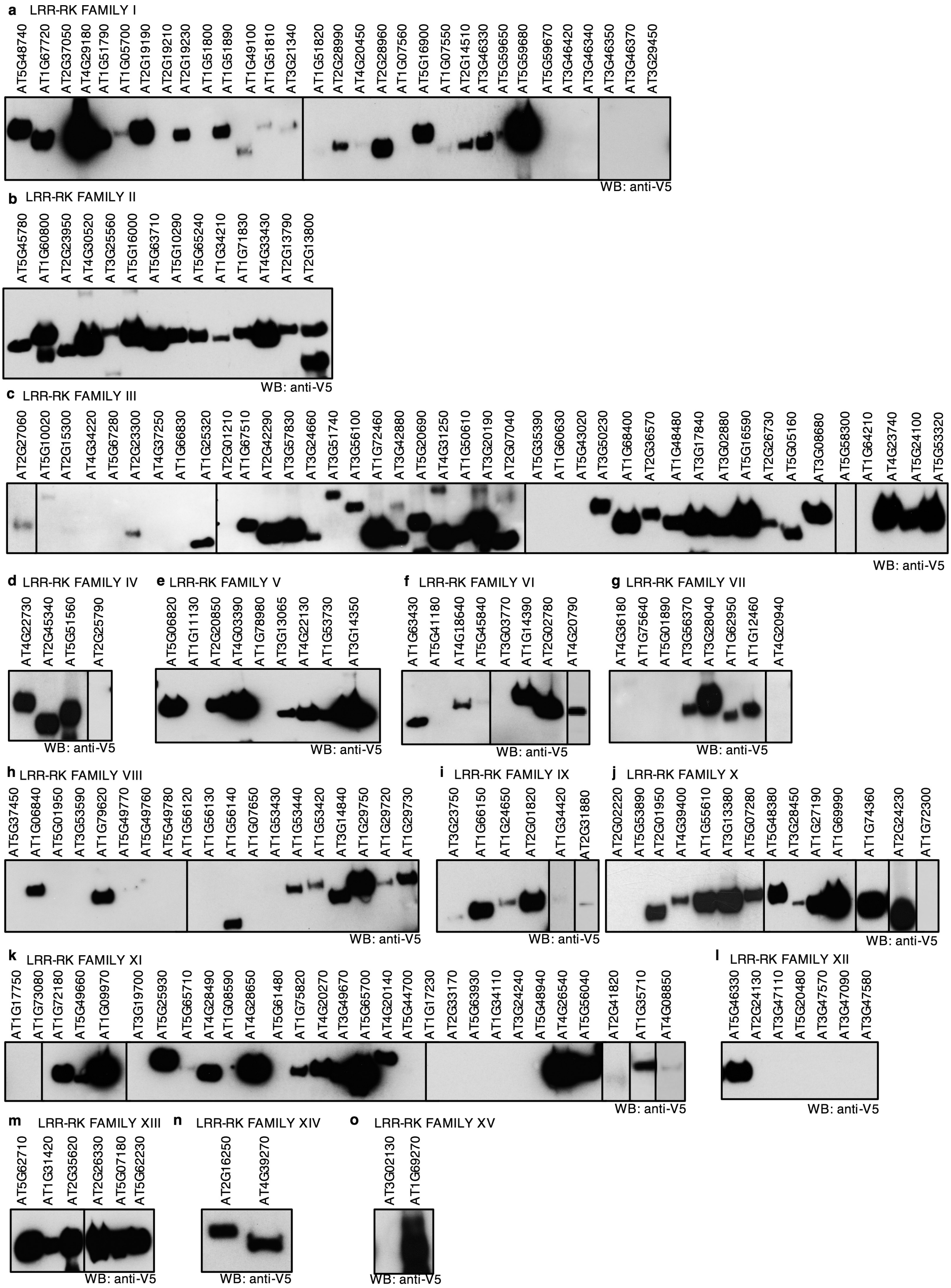
To estimate the reliability of the estimates provided by the retest screen (Extended Data Fig. 3d), the observed rate of interactions found in the CSI and retest sets were used for a Monte Carlo simulation. Sets of observations were selected at random from these populations, with the number of observations equal to the number present in the retest sets. This process was completed 100,000 times. These values were used to calculate the mean and s.d. of the samplings.

Details about the linear mixed effect modelling can be found in the Supplementary Methods.

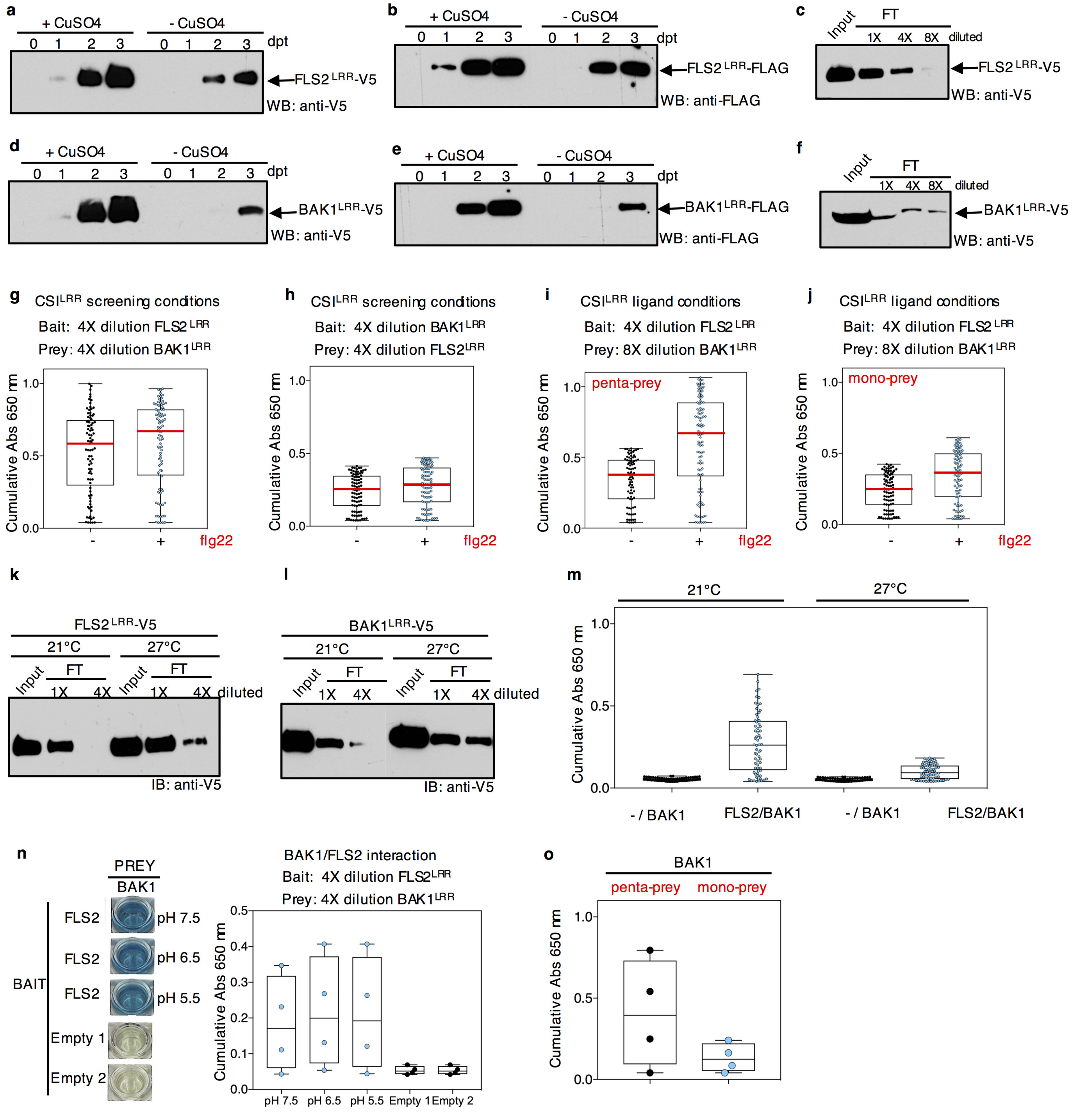
Data and software accessibility. The data supporting the findings of this study are available within the paper and its Supplementary Information files. Source data for Figs 1, 2, 3, 4, Extended Data Fig. 2, 3, 4, 5, 6, 8 and 9 are provided with the paper. All the raw absorbance reads related to the ECD interaction screen are available in the Supplementary Table 11. The high-confidence LRR-RKs interaction dataset is publically available online at the Bio-Analytic Resource under accession (MI 2189, Smakowska-Luzan et al. 2018, doi:10.1038/nature25184): http://bar.utoronto.ca/interactions. The custom PLATERO script used for concatenating the interaction absorbance values is available upon request from the corresponding author or from https://github.com/AdamMott/platero-code.

26. Couto, D. & Zipfel, C. Regulation of pattern recognition receptor signalling in plants. *Nat. Rev. Immunol.* **16,** 537–552 (2016).
27. McWilliam, H. *et al.* Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.* **41,** W597–W600 (2013).
28. Gou, X. *et al.* Genome-wide cloning and sequence analysis of leucine-rich repeat receptor-like protein kinase genes in *Arabidopsis thaliana*. *BMC Genomics* **11,** 19 (2010).
29. Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J. & Nadon, R. Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* **24,** 167–175 (2006).
30. Brideau, C., Gunter, B., Pikounis, B. & Liaw, A. Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* **8,** 634–647 (2003).
31. Halter, T. *et al.* The leucine-rich repeat receptor kinase BIR2 is a negative regulator of BAK1 in plant immunity. *Curr. Biol.* **24,** 134–143 (2014).
32. Eyüboglu, B. *et al.* Molecular characterisation of the STRUBBELIG-RECEPTOR FAMILY of genes encoding putative leucine-rich repeat receptor-like kinases in *Arabidopsis thaliana*. *BMC Plant Biol.* **7,** 16 (2007).
33. Jordá, L. *et al.* ERECTA and BAK1 receptor like kinases interact to regulate immune responses in *Arabidopsis*. *Front. Plant Sci.* **7,** 897 (2016).
34. Russinova, E. *et al.* Heterodimerization and endocytosis of *Arabidopsis* brassinosteroid receptors BRI1 and AtSERK3 (BAK1). *Plant Cell* **16,** 3216–3229 (2004).
35. Roux, M. *et al.* The *Arabidopsis* leucine-rich repeat receptor-like kinases BAK1/SERK3 and BKK1/SERK4 are required for innate immunity to hemibiotrophic and biotrophic pathogens. *Plant Cell* **23,** 2440–2455 (2011).
36. Hazak, O. *et al.* Perception of root-active CLE peptides requires CORYNE function in the phloem vasculature. *EMBO Rep.* **18,** 1367–1381 (2017).
37. Meng, X. *et al.* Differential function of *Arabidopsis* SERK family receptor-like kinases in stomatal patterning. *Curr. Biol.* **25,** 2361–2372 (2015).
38. Nodine, M. D., Yadegari, R. & Tax, F. E. RPK1 and TOAD2 are two receptor-like kinases redundantly required for *Arabidopsis* embryonic pattern formation. *Dev. Cell* **12,** 943–956 (2007).
39. Wang, X. *et al.* IDL6-HAE/HSL2 impacts pectin degradation and resistance to *Pseudomonas syringae* pv *tomato* DC3000 in *Arabidopsis* leaves. *Plant J.* **89,** 250–263 (2017).
40. Meng, X. *et al.* Ligand-induced receptor-like kinase complex regulates floral organ abscission in *Arabidopsis*. *Cell Reports* **14,** 1330–1338 (2016).
41. Santiago, J. *et al.* Mechanistic insight into a peptide hormone signaling complex mediating floral organ abscission. *Elife* **5,** e15075 (2016).
42. Nimchuk, Z. L. CLAVATA1 controls distinct signaling outputs that buffer shoot stem cell proliferation through a two-step transcriptional compensation loop. *PLoS Genet.* **13,** e1006681 (2017).
43. Cole, S. J. & Diener, A. C. Diversity in receptor-like kinase genes is a major determinant of quantitative resistance to *Fusarium oxysporum* f.sp. *matthioli*. *New Phytol.* **200,** 172–184 (2013).
44. Agusti, J., Lichtenberger, R., Schwarz, M., Nehlin, L. & Greb, T. Characterization of transcriptome remodeling during cambium formation identifies MOL1 and RUL1 as opposing regulators of secondary growth. *PLoS Genet.* **7,** e1001312 (2011).

45. Xiao, D. *et al.* SENESCENCE-SUPPRESSED PROTEIN PHOSPHATASE directly interacts with the cytoplasmic domain of SENESCENCE-ASSOCIATED RECEPTOR-LIKE KINASE and negatively regulates leaf senescence in *Arabidopsis*. *Plant Physiol.* **169,** 1275–1291 (2015).
46. Kang, Y. H. & Hardtke, C. S. *Arabidopsis* MAKR5 is a positive effector of BAM3-dependent CLE45 signaling. *EMBO Rep.* **17,** 1145–1154 (2016).
47. Albert, I. *et al.* An RLP23–SOBIR1–BAK1 complex mediates NLP-triggered immunity. *Nat. Plants* **1,** 15140 (2015).
48. Valon, C., Smalle, J., Goodman, H. M. & Giraudat, J. Characterization of an *Arabidopsis thaliana* gene (*TMKL1*) encoding a putative transmembrane protein with an unusual kinase-like domain. *Plant Mol. Biol.* **23,** 415–421 (1993).
49. Tarutani, Y. *et al.* Molecular characterization of two highly homologous receptor-like kinase genes, *RLK902* and *RKL1*, in *Arabidopsis thaliana*. *Biosci. Biotechnol. Biochem.* **68,** 1935–1941 (2004).
50. Chang, C. *et al.* The *TMK1* gene from *Arabidopsis* codes for a protein with structural and biochemical characteristics of a receptor protein kinase. *Plant Cell* **4,** 1263–1271 (1992).
51. Wang, T. *et al.* A receptor heteromer mediates the male perception of female attractants in plants. *Nature* **531,** 241–244 (2016).
52. Sun, W. *et al.* Probing the *Arabidopsis* flagellin receptor: FLS2–FLS2 association and the contributions of specific domains to signaling function. *Plant Cell* **24,** 1096–1113 (2012).
53. Yeh, Y. H. & Panzeri, D. The *Arabidopsis* malectin-like/LRR-RLK IOS1 is critical for BAK1-dependent and BAK1-independent pattern-triggered immunity. *Plant Cell* **28,** 1701–1721(2016).
54. Igarashi, D., Tsuda, K. & Katagiri, F. The peptide growth factor, phytosulfokine, attenuates pattern-triggered immunity. *Plant J*. **71,** 194–204 (2012).
55. Mosher, S. *et al.* The tyrosine-sulfated peptide receptors PSKR1 and PSY1R modify the immunity of *Arabidopsis* to biotrophic and necrotrophic pathogens in an antagonistic manner. *Plant J.* **73,** 469–482 (2013).
56. Wang, J. *et al.* Allosteric receptor activation by the plant peptide hormone phytosulfokine. *Nature* **525,** 265–268 (2015).
57. Postel, S. *et al.* The multifunctional leucine-rich repeat receptor kinase BAK1 is implicated in *Arabidopsis* development and immunity. *Eur. J. Cell Biol.* **89,** 169–174 (2010).
58. Sakamoto, T. *et al.* The tomato RLK superfamily: phylogeny and functional predictions about the role of the LRRII-RLK subfamily in antiviral defense. *BMC Plant Biol*. **12,** 229 (2012).

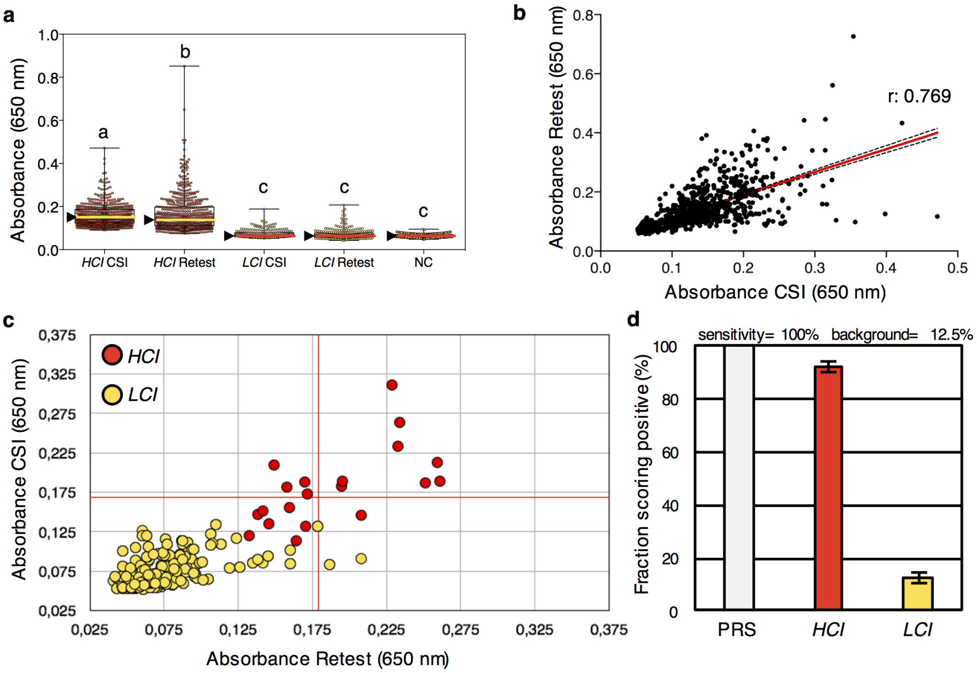


Extended Data Figure 1 | Expression profiles of LRR-RK ECDs produced as recombinant baits with the *Drosophila* S2 cells protein expression system. a–o, Western blot analyses of raw supernatants from S2 cells transfected with ECD expression vectors. Blots were cropped and arranged to match the phylogenetic tree of the LRR-RK gene family. The family subclasses and *Arabidopsis* gene initiative (AGI) identifiers are indicated at the top. For lanes showing no obvious anti-V5 signals, a mild concentration of the S2 cell media and/or purification on protein-A-coated 96-well plates allowed for confirmation of expression and secretion of the ECDs. This experiment was conducted once with the full set of 200 ECDs. The expression of 130 independently expressed ECDs was tested one additional time with similar results.

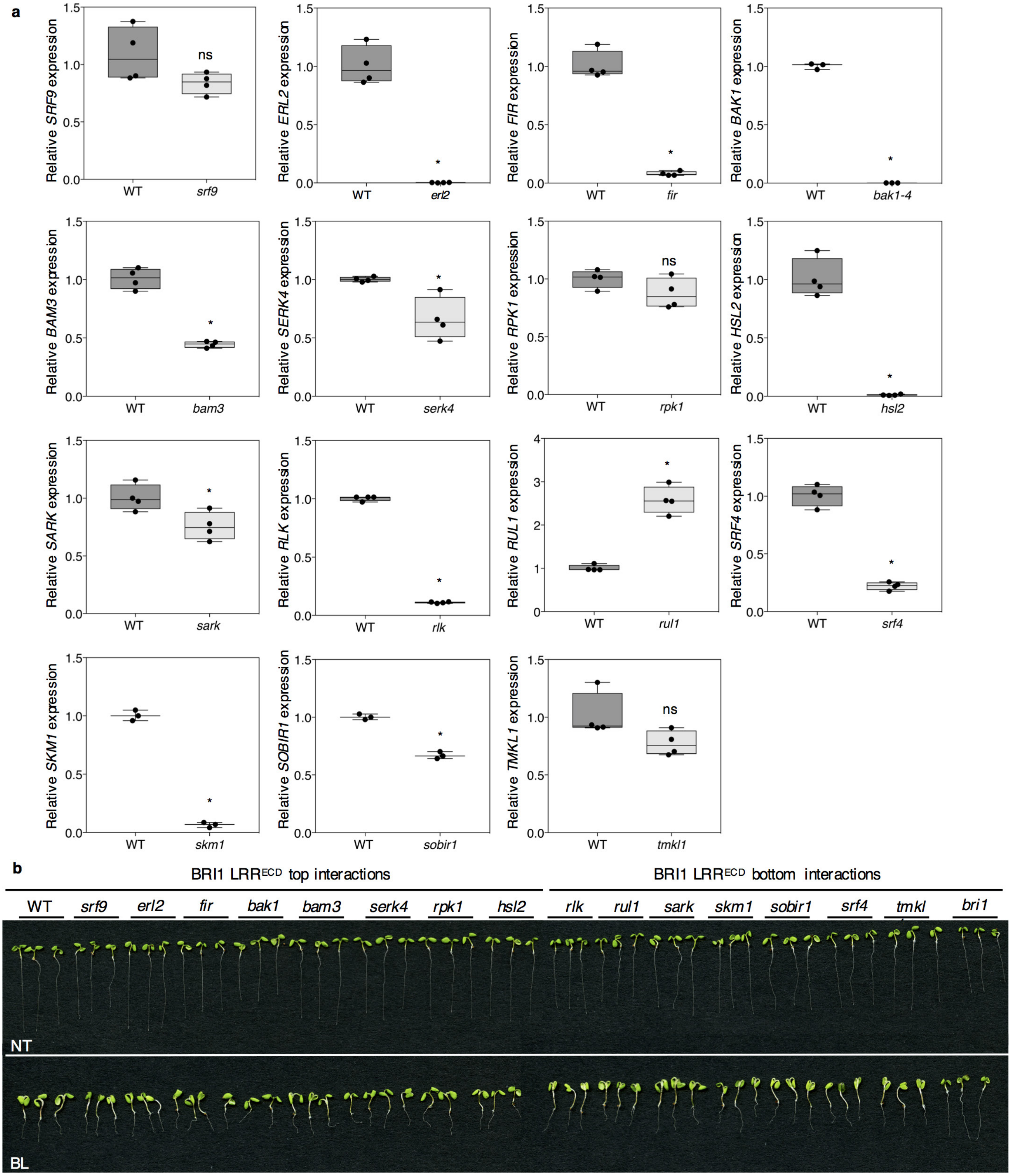


Extended Data Figure 2 | See next page for caption.

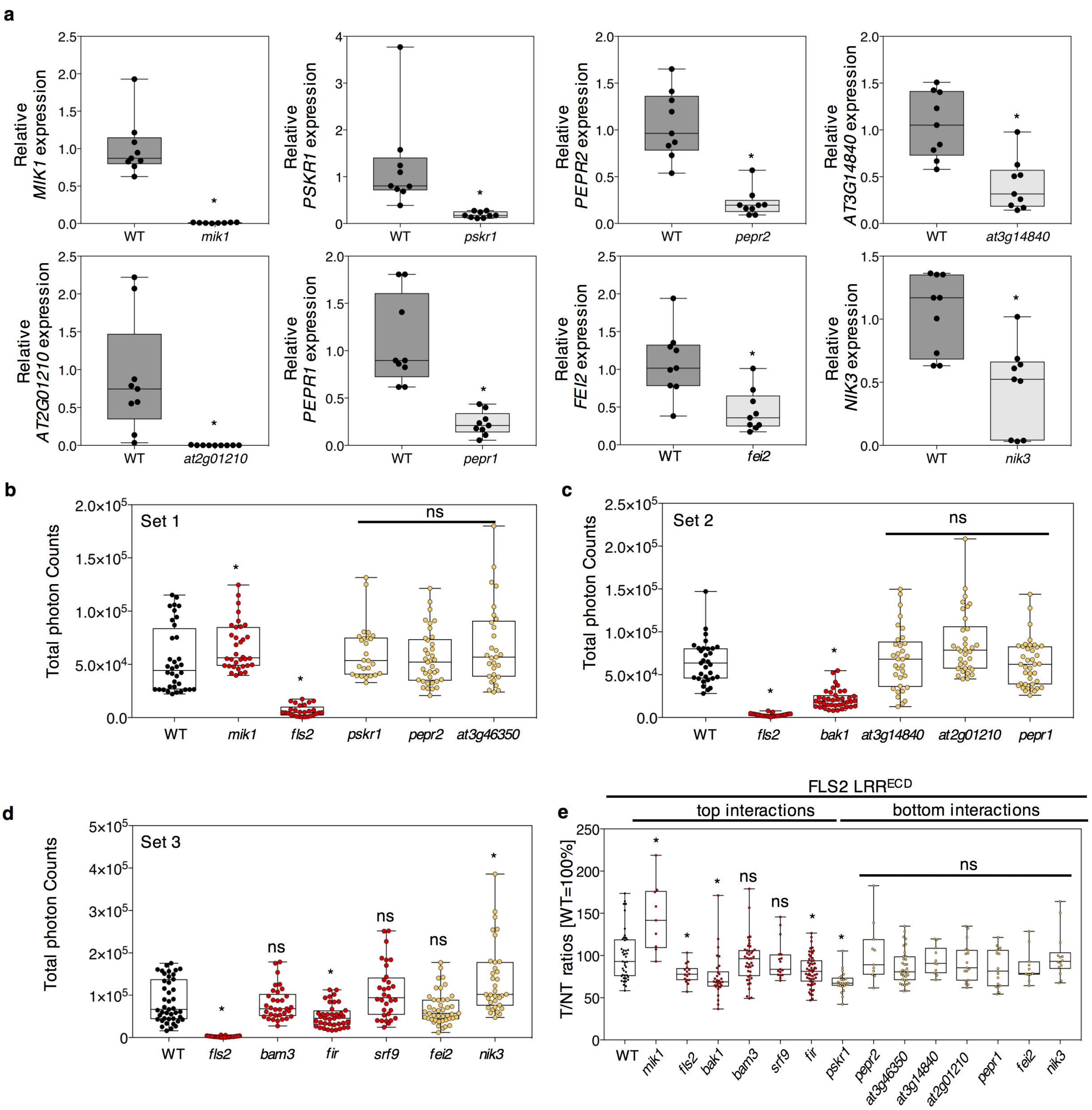
Extended Data Figure 2 | Calibration of the CSI^{LRR} screen conditions on ligand-dependent (FLS2–BAK1) and ligand-independent (BAK1–BIR4) interaction pairs. a, **b**, Western blot analyses of raw supernatants from S2 cells transfected with prey and bait expression vectors for the ECD of FLS2 (bait, western blot: anti-V5 antibody; prey, western blot: anti-Flag antibody). S2 cells left untreated (−) or treated with $CuSO_4$ (+). Days post transfection (dpt) are indicated on top. The experiment was repeated independently twice with similar results. **c**, Binding of the FLS2 ECD to the protein-A-coated 96-well plates. A fourfold dilution (4×) of the insect cell medium containing the ECD of FLS2 saturates the binding sites of protein-A-coated wells as indicated by immunoblots of the flow-through (FT). The experiment was repeated independently twice with similar results. **d**–**f**, As in **a**–**c** but for BAK1. The experiment was repeated independently twice with similar results. **g**, Plate interaction assays between the ECDs of BAK1 (prey) and FLS2 (bait) represented as cumulative absorbance (Abs 650 nm) over 18 h. Dots represent individual observations at each hour from five technical replicates. Box plots display the first and third quartiles, split by the median (red line); whiskers extend to include the maximum and minimum values. The presence of flg22 (+) in fourfold-diluted CSI^{LRR} screening conditions weakly promotes the interaction between the two ECDs. **h**, Technical replicates and box plots are as in **g**, but with BAK1 (bait) and FLS2 (prey). **i**, Technical replicates and box plots are as in **g** but with BAK1 (prey eightfold diluted) and FLS2 (bait fourfold diluted). In these conditions, the binding between the ECDs of BAK1 and FLS2 is largely enhanced by the presence of flg22 (+), indicating that the proteins produced in our expression system can interact in a ligand-dependent manner and are thus functional. **j**, Technical replicates and box plots as in **g**, but using a prey variant of BAK1 that can no longer pentamerize owing to the deletion of the COMP domain (BAK1 mono-prey). Binding between the two ECDs is still observed, but at a reduced level, thus indicating the importance of the pentamerization motif for detecting transient and low affinity interactions in the absence of ligand. **k**, **l**, Binding of FLS2 and BAK1 ECDs to protein-A-coated 96-well plates (as indicated by immunoblots of the flow-through) when proteins are produced from S2 cells growing either at 21 °C or 27 °C. Immunoblots show a slight increase in protein production at 27 °C with similar binding capacities to the protein-A-coated plate. The protein expression levels at the two temperatures were assessed more than three times with similar results. The plate saturation experiment for proteins produced at 27 °C was conducted once. **m**, Plate interaction assays between BAK1 (prey) and FLS2 (bait) (in fourfold-diluted conditions) represented as cumulative absorbance (Abs 650 nm) over a 150-min time course. Dots represent individual observations made every 10 min from four technical replicates. Box plots as in **g**. Although slightly more abundant, proteins produced at 27 °C do not interact as well when produced at 21 °C. Protein expression for the CSI^{LRR} screen was performed at 21 °C. **n**, The FLS2–BAK1 interaction is insensitive to changes in pH conditions. Left, the interaction between FLS2 (bait) and BAK1 (prey) was observed in the pH range from 5.5 to 7.5. This experiment was conducted once. Right, plate interaction assays between BAK1 (prey) and FLS2 (bait) (in fourfold-diluted conditions) represented as cumulative absorbance over a 3-h time course. Dots represent individual observations at each hour from one technical replicate. The CSI^{LRR} screen was performed at the pH of the conditioned S2 cells supernatant (~pH 7.5). **o**, Plate interaction assays between BAK1 (as mono-prey (blue dots) or penta-prey (black dots)) and BIR4 represented as cumulative absorbance over a 3-h time course. Dots represent individual observations at each hour from one technical replicate. This experiment was conducted once. The data indicate that the pentamerization of the prey is a key requirement for enhancing the interaction detection sensitivity, without disrupting the functionality of the ECDs. BAK1 and BIR4 are ligand-independent interaction partners and the screening conditions used are also appropriate to detect this interaction.



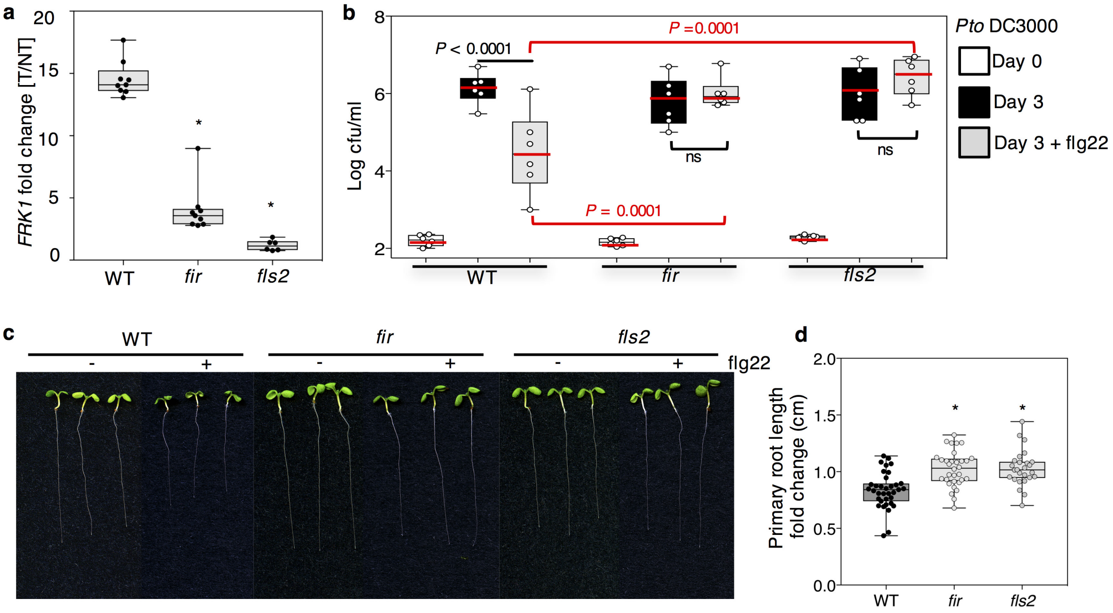
Extended Data Figure 3 | Comparison of the primary and retest screens parameters. a, Geometric mean of the normalized absorbance values for the HCI (red dots) and LCI (yellow dots) obtained from the primary screen (CSI), the validation screen (retest) and the negative controls (NC) associated with the two screens. n denotes numbers of bidirectional interactions: HCI CSI ($n = 567$), HCI retest ($n = 567$), LCI CSI ($n = 248$), LCI retest ($n = 248$), and NC ($n = 618$). The box plots contain the first and third quartiles, split by the median (yellow or red lines indicated by the arrow on the left of the boxes); whiskers extend to include the maximum and minimum values. Statistical significance was determined using unbalanced one-way ANOVA by Tukey's HSD for all pairwise comparisons. Datasets with the same letter are indistinguishable at >95% confidence. **b**, Plots of a linear regression for the entire set of normalized absorbance values obtained from the retest screens (absorbance retest; y axis) and the corresponding values from the from the primary screen (absorbance CSI; x axis). The thick, straight red line is the linear regression that best describes the entire set of data points (Spearman's $r = 0.7696$; indicated on top). The fine red dashed lines represent the 95% confidence intervals of the regression. $n = 815$ bidirectional interactions. **c**, Comparison of the geometric mean of normalized absorbance values for selected interactions. Values from the primary screen (absorbance CSI; y axis) and the validation screen (absorbance retest; x axis) are shown for the LCI set (yellow dots) and for 20 interactions selected at random from the HCI set (red dots). The number of interactions shown for each set was selected to approach the numbers present in the entire interaction search space. The red lines show the absorbance values corresponding to the FLS2–BAK1 interaction in both screens. **d**, Retest assay performance parameters interpreted within the performance window measured by positive reference set (PRS) and LCI calibration. To estimate the reliability of the estimates provided by the retest, the observed rate of interactions found in the HCI and LCI sets were used for a Monte Carlo simulation. $n = 100{,}000$ independent sets of observations selected at random from these populations, with the number of observations equal to the number present in the retest sets. These values were used to calculate the mean and s.d. of the samplings, which are presented as error bars.



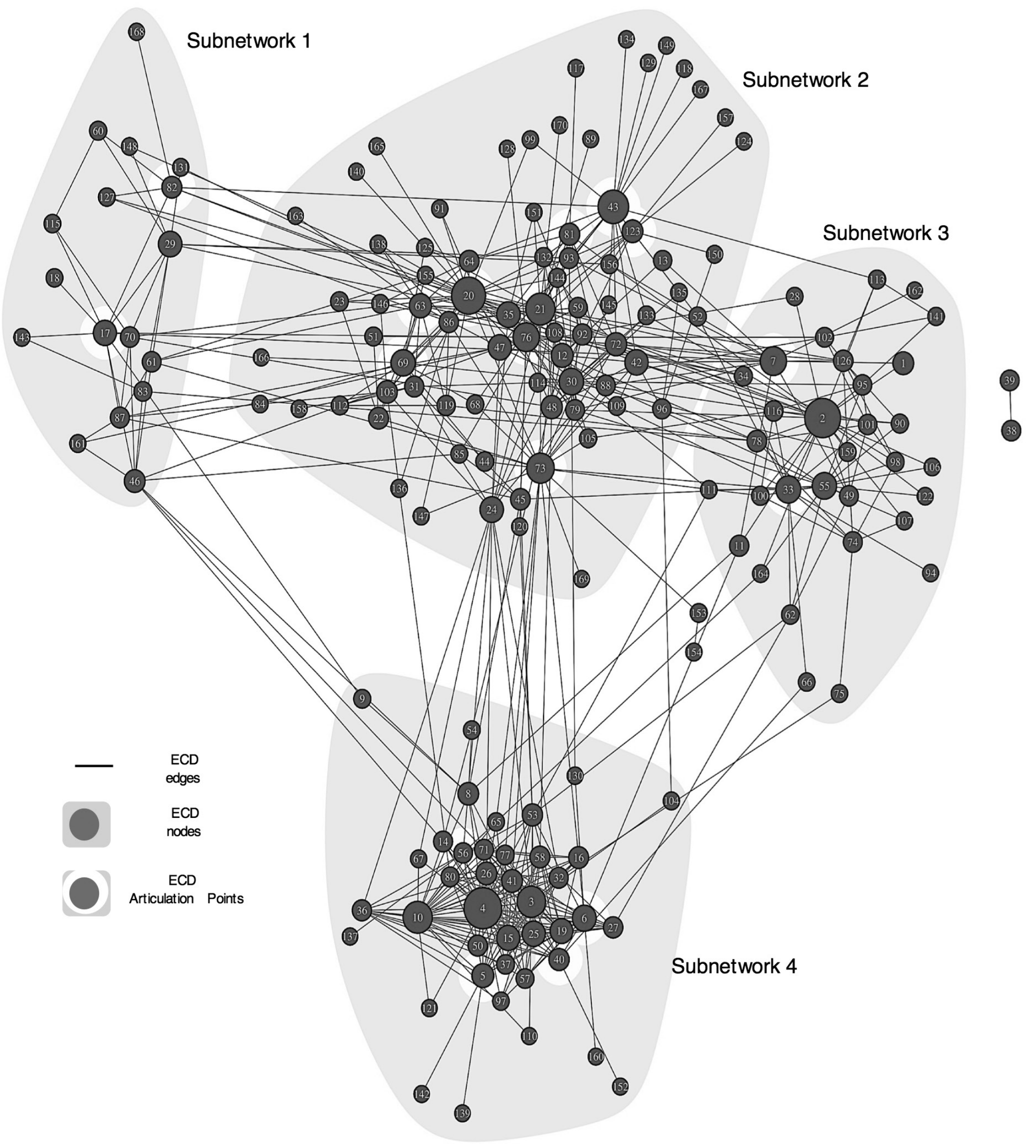
Extended Data Figure 4 | Characterization of BRI1 interaction partners. a, qPCR analyses showing altered gene expression in T-DNA lines targeting the interaction partners of BRI1 (Fig. 1b). Genotypes are indicated. Relative expression levels were calculated and *ACTIN* was used as reference gene to control for cDNA amount in each reaction. The box plots contain the first and third quartiles, split by the median; whiskers extend to include the maximum and minimum values. $n = 4$ biologically independent mRNA samples for all genotypes, except for *bak1-4*, *skm1* and *sobir1* where $n = 3$. Statistical significance was estimated by an unpaired two-sided *t*-test and is indicated on top of the boxes: *erl2* $*P = 0.0012$, *fir* $*P = 5.3508 \times 10^{-6}$, *bak1-4* $*P = 3.08212 \times 10^{-7}$, *bam3* $*P = 1.9378 \times 10^{-5}$, *serk4* $*P = 0.0108$, *hsl2* $*P = 2.06945 \times 10^{-5}$, *sark* $*P = 0.0259$, *rlk* $*P = 2.12971 \times 10^{-10}$, *rul1* $*P = 7.49918 \times 10^{-5}$, *srf4* $*P = 3.08212 \times 10^{-7}$, *skm1* $*P = 5.5911 \times 10^{-6}$, *sobir1* $*P = 0.0001$. ns, not significant. **b**, T-DNA insertions targeting the HCI (top interactions) and LCI (bottom interaction) partners of BRI1. Morphology of representative seedlings grown for 7 days in the absence (NT) or presence (BL) of 500 nM brassinolide, the most potent brassinosteroid. Genotypes are indicated. The experiment was conducted six times with similar results.



Extended Data Figure 5 | Characterization of FLS2 interaction partners. a, qPCR analyses showing altered gene expression in T-DNA lines targeting the interaction partners of FLS2 (Fig. 1c). Genotypes are indicated. Relative expression levels were calculated and *ACTIN* was used as reference gene to control for cDNA amount in each reaction. $n=9$ biologically independent mRNA samples for all tested genotypes. Statistical significance was estimated by an unpaired two-sided *t*-test: *mik1* $*P=8.17192\times10^{-6}$, *pskr1* $*P=0.007$, *pepr2* $*P=0.007$, *at3g14840* $*P=0.005$, *at2g01210* $*P=0.0032$, *pepr1* $*P=1.16519\times10^{-5}$, *fei2* $*P=0.005$, *nik3* $*P=0.0015$. **b**, Oxidative burst represented as total photon counts, triggered by 1 μM flg22 in wild type (black) and mutant lines targeting the HCI (top; red) and LCI (bottom, yellow) partners for FLS2. Genotypes are indicated. Dots represent individual observations from four independent experiments. *n* denotes numbers of biologically independent leaf discs: WT ($n=36$), *mik1* ($n=36$), *fls2* ($n=28$), *pskr1* ($n=27$), *pepr2* ($n=38$), *at3g46350* ($n=39$). Statistical significance was determined using linear mixed effect modelling, and symbols indicate the results of a post hoc unpaired two-sided *t*-test corrected with the Holm method for multiple testing: *mik1* $*P=4.32\times10^{-2}$, *fls2* $*P=1\times10^{-15}$. **c**, As in **b**, except: WT ($n=32$), *fls2* ($n=27$), *bak1* ($n=39$), *at3g14840* ($n=33$), *at2g01210* ($n=38$), *pepr1* ($n=40$). *bak1* $*P=1\times10^{-15}$, *fls2* $*P=1\times10^{-15}$. **d**, As in **b** and **c**, except: WT ($n=43$), *fls2* ($n=29$), *bam3* ($n=33$), *fir* ($n=39$), *srf9* ($n=32$), *fei2* ($n=45$), *nik3* ($n=32$). *fir* $*P=1.38\times10^{-3}$, *fls2* $*P=1.2\times10^{-15}$, *nik3* $*P=1.38\times10^{-3}$. The ROS burst assays in **b**–**d** were performed on independent plates (set number) and every plate contained wild type and *fls2* controls, as well as randomly assigned mutant lines. **e**, flg22-induced peroxidase (POX) assay in wild-type (black bar) and mutant lines targeting the HCI (top interactions; red) and LCI (bottom interactions, yellow) partners for FLS2. Genotypes are indicated. Leaf disks from 4-week-old plants were treated with water (NT) or 1 μM flg22 (T). The level of flg22-induced POX was normalized to the corresponding non-treated control. The level of POX present in the wild type was set to 100 for easier interpretation. *n* denotes numbers of biologically independent leaf discs from two independent experiments: WT ($n=44$), *mik1* ($n=10$), *fls2* ($n=17$), *bak1* ($n=31$), *bam3* ($n=42$), *srf9* ($n=18$), *fir* ($n=55$), *pskr1* ($n=24$), *pepr2* ($n=12$), *at3g46350* ($n=36$), *at3g14840* ($n=12$), *at2g01210* ($n=18$), *pepr1* ($n=12$), *fei2* ($n=11$), *nik3* ($n=15$). Statistical significance was estimated using a paired two-sided *t*-test for each genotype, corrected for multiple tests using the Holm–Bonferroni correction. *mik1* $*P=5.71\times10^{-4}$, *fls2* $*P=0.046$, *bak1* $*P=0.0039$, *fir* $*P=0.0048$, *pskr1* $*P=9.49\times10^{-5}$. All box plots contain the first and third quartiles, split by the median; whiskers extend to include the maximum and minimum values.

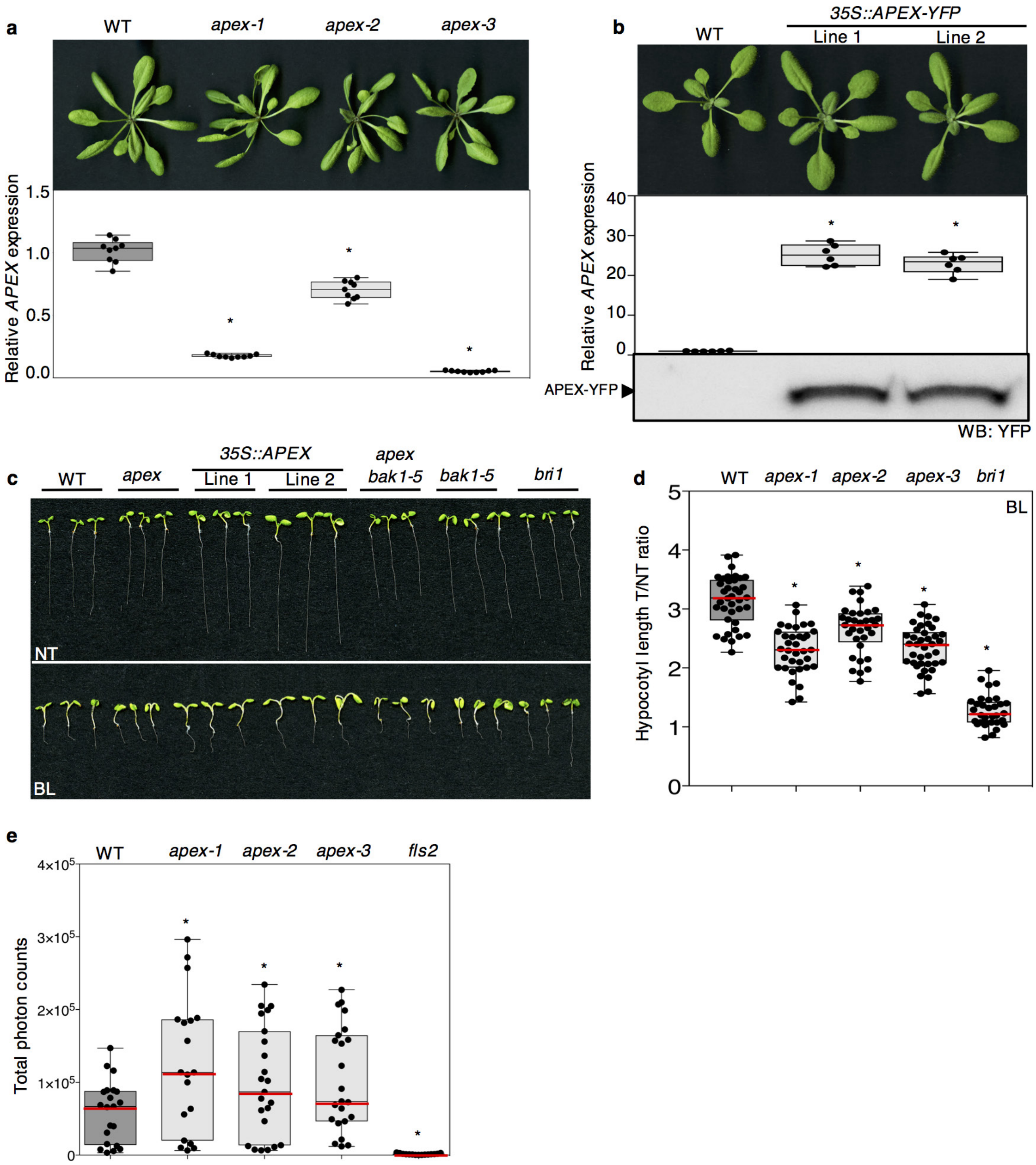


Extended Data Figure 6 | FIR regulates flg22-induced responses. **a**, Seedlings of the genotypes indicated on the bottom were treated with either water (NT) or flg22 (T) and changes in *FRK1* transcript levels were quantified by qPCR analyses. Dots represent individual observations from three independent experiments. *n* denotes numbers of biologically independent mRNA samples: WT ($n = 9$ (NT), $n = 9$ (T)), *fir* ($n = 9$, $n = 9$) and *fls2* ($n = 6$, $n = 6$). Statistical significance was determined using linear mixed effect modelling followed by comparison of each genotype to the wild-type control using unpaired two-sided *t*-test followed by multiple testing correction using the Holm method. *fir* $*P = 1.42 \times 10^{-7}$, *fls2* $*P = 4 \times 10^{-16}$. **b**, Growth of *Pto* DC3000 on the genetic backgrounds indicated at the bottom of the chart. Four-week-old plants were infiltrated with 10^5 cfu ml^{-1} in the absence (black bars) or presence (grey bars) of 1 μM flg22. The number of bacteria per area of leaf (cfu ml^{-1}) was plotted on a $\log_{10}$ scale for day 0 (open bars) and day 3 (closed bars). Dots represent individual observations from two independent experiments. *n* denotes numbers of samples, each including 4 biologically independent leaf discs. For day 0, WT ($n = 6$), *fir* ($n = 6$), *fls2* ($n = 6$); for day 3, WT ($n = 6$), *fir* ($n = 6$), *fls2* ($n = 6$); for day 3 + flg22, WT ($n = 6$), *fir* ($n = 6$), *fls2* ($n = 6$). Statistical significance for bacterial growth was estimated by two-way ANOVA. A third experiment performed at an inoculum of 10^6 cfu ml^{-1} corroborated these results. **c**, Morphology of 7-day-old seedlings grown in the absence (−) or presence (+) of 1 μM flg22. Genotypes are indicated. The experiment was conducted twice with similar results. **d**, Primary root length (cm) from seedlings grown in the presence (T) or absence (NT) of 1 μM flg22. Fold changes are T/NT ratios. Dots represent individual observations from two independent experiments. *n* denotes the following numbers of biologically independent roots: WT ($n = 32$ (NT), $n = 36$ (T)), *fir* ($n = 34$ (NT), $n = 32$ (T)), *fls2* ($n = 27$ (NT), $n = 26$ (T)). Statistical significance for two biological replicates was determined using linear mixed effect modelling followed by comparison of each genotype to the wild-type control using unpaired two-sided *t*-test followed by multiple testing correction using the Holm method. *fir* $*P = 2.02 \times 10^{-6}$, *fls2* $*P = 2.02 \times 10^{-6}$. All box plots display the first and third quartiles, split by the median; whiskers extend to include the maximum and minimum values.



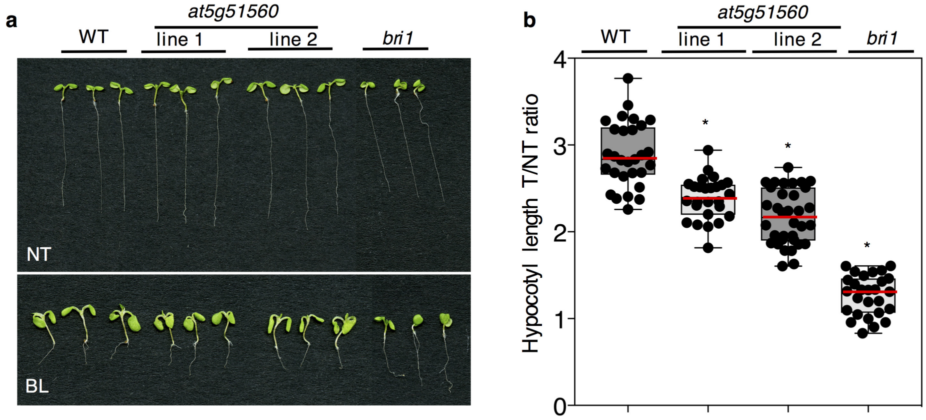
- 1 RGI5
- 2 RLK902
- 3 SARK
- 4 SERK3/BAK1
- 5 BIR4
- 6 NIK1
- 7 PRK4
- 8 FLS2
- 9 MIK1/PXL2
- 10 ERL2
- 11 AT3G53590
- 12 AT4G23740
- 13 PXY
- 14 BRI1
- 15 SRF9
- 16 BIR3
- 17 AT5G49770
- 18 BARK1
- 19 RPK1
- 20 AT2G02780
- 21 NIK3
- 22 AT1G51820
- 23 AT2G42290
- 24 GSO1
- 25 AT2G27060
- 26 ERECTA
- 27 ERL1
- 28 IKU2
- 29 AT1G17230
- 30 AT1G68400
- 31 AT3G28040
- 32 SRF3
- 33 AT1G49100
- 34 AT4G36180
- 35 TMK1
- 36 SERK5
- 37 HSL2
- 38 AT2G24130
- 39 AT5G41180
- 40 BRL1
- 41 IOS1
- 42 AT1G63430
- 43 AT5G63710
- 44 AT1G67720
- 45 AT5G62710
- 46 BAM3
- 47 RHS16
- 48 AT5G45780
- 49 AT1G67510
- 50 SERK4
- 51 CLV1
- 52 AT5G25930
- 53 BRL3
- 54 RPK2
- 55 AT3G50230
- 56 SRF7
- 57 PSY1R
- 58 SRF6
- 59 PRK7
- 60 AT1G07650
- 61 AT1G24650
- 62 BAM1
- 63 AT2G28960
- 64 AT4G37250
- 65 SRF5
- 66 AT5G58300
- 67 SRF1
- 68 AT3G47090
- 69 PXC2
- 70 AT3G14840
- 71 BIR1
- 72 PSKR2
- 73 AT5G51560
- 74 AT5G53320
- 75 PRK1
- 76 AT1G64210
- 77 AT2G14510
- 78 AT1G51790
- 79 SOBIR1
- 80 SERK1
- 81 AT5G24100
- 82 RLK7/KUK2
- 83 BRL2
- 84 AT1G07550
- 85 AT5G10290
- 86 SERK2
- 87 RKL1
- 88 PSKR1
- 89 AT5G48740
- 90 AT1G56140
- 91 AT4G39270
- 92 PEPR2
- 93 AT2G01820
- 94 AT2G16250
- 95 AT1G51890
- 96 AT2G23950
- 97 EFR
- 98 AT1G56120
- 99 AT5G37450
- 100 AT3G21340
- 101 AT1G72460
- 102 AT4G29450
- 103 AT1G25320
- 104 SRF8
- 105 AT3G46420
- 106 AT2G01210
- 107 AT2G19230
- 108 PRK5
- 109 LRR1
- 110 SRF2
- 111 BAM2
- 112 FRK1/SIRK
- 113 MEE39
- 114 AT3G57830
- 115 AT1G35710
- 116 AT2G45340
- 117 MDIS2/MRH1
- 118 PEPR1
- 119 AT1G12460
- 120 AT1G08590
- 121 NIK2
- 122 AT1G53430
- 123 AT1G14390
- 124 RKF1
- 125 AT4G20450
- 126 AT3G02880
- 127 TMKL1
- 128 EMS1
- 129 AT3G03770
- 130 AT1G06840
- 131 AT2G24230
- 132 RLK
- 133 AT1G62950
- 134 AT3G46340
- 135 AT2G23300
- 136 AT5G59680
- 137 AT4G22730
- 138 AT5G10020
- 139 RGFR3
- 140 SKM1
- 141 AT3G08680
- 142 AT1G07560
- 143 RGFR1
- 144 AT1G79620
- 145 AT4G20790
- 146 AT2G28990
- 147 AT5G01950
- 148 FEI1
- 149 AT3G56370
- 150 CEPR2
- 151 MDIS1
- 152 XIP1/CEPR1
- 153 AT3G47580
- 154 MRLK
- 155 AT1G51810
- 156 AT1G74360
- 157 HAESA
- 158 AT2G37050
- 159 AT3G46350
- 160 AT2G19210
- 161 SKM2
- 162 RGFR2
- 163 GHR1
- 164 AT1G29730
- 165 FEI2
- 166 AT1G05700
- 167 AT5G16900
- 168 AT5G63930
- 169 AT5G59650
- 170 PRK6

Extended Data Figure 7 | CSI^{LRR} network representation and table of nodes with their corresponding identification numbers or acronyms. The network construction and other features are the same as shown in Fig. 2b. The nodes surrounded by white halos are articulation points. The numbers in each node corresponding to the ECD of each LRR-RK are shown in the table.



Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Characterization of independent *apex* mutant and *35S::APEX* transgenic lines. **a**, Top, rosette morphology of 4-week-old wild-type, *apex-1* and *apex-2*, and *apex-3* knockdown lines grown under long-day photoperiod at 22 °C. Genetic backgrounds are indicated. No obvious changes in rosette morphology are observed. The experiment was conducted three times with similar results. Bottom, qPCR analyses showing fold reduction of *APEX* transcripts in the independent mutant lines. Relative expression levels were calculated and *ACTIN* was used as reference control gene. Dots represent individual observations from three independent experiments. $n = 9$ biologically independent mRNA samples for each genotype. Statistical significance was determined using linear mixed effect modelling followed by comparison of each genotype to the wild-type control using unpaired two-sided t-test followed by multiple testing correction using the Holm method. *apex-1* $^*P = 6 \times 10^{-16}$, *apex-2* $^*P = 5.33 \times 10^{-15}$, *apex-3* $^*P = 6 \times 10^{-16}$. **b**, Top, rosette morphology of 3-week-old wild type and *35S::APEX* lines 1 and 2 grown under long-day photoperiod at 22 °C. Genetic backgrounds are indicated on the top. Rosettes of *35S::APEX* lines are slightly larger than WT under long-day photoperiod at 22 °C. The experiment was conducted three times with similar results. Middle: Quantitative real-time PCR analyses showing fold induction of the *APEX* transgene in the overexpression lines used in this study. Relative expression levels were calculated and *ACTIN* was used as reference gene to control for cDNA amount in each reaction. Dots represent individual observations from two independent experiments. $n = 6$ biologically independent mRNA samples for each genotype. Statistical significance was determined using linear mixed effect modelling followed by comparison of each genotype to the WT control using an unpaired two-sided t-test followed by multiple testing correction using the Holm method and is indicated on top of the boxes: *35S::APEX* line 1 $^*P = 3.38 \times 10^{-14}$, *35S::APEX* line 2 $^*P = 7.77 \times 10^{-14}$. Bottom, detection of APEX–YFP in stable transgenic T_3 lines by western blot using an anti-GFP antibody. **c**, Modulation of BRI1 signalling by APEX gene dosage. Morphology of representative seedlings corresponding to Fig. 4a. Genotypes are indicated. The experiment was conducted over three times with similar results. **d**, Hypocotyl length ratios of seedlings grown in the presence (T) or absence (NT) of 500 nM brassinolide (BL). Genotypes are indicated. Dots represent individual observations from three independent experiments. n denotes numbers of biologically independent hypocotyls. WT ($n = 43$ (NT), $n = 33$ (T)), *apex-1* ($n = 31$, $n = 35$), *apex-2* ($n = 32$, $n = 33$), *apex-3* ($n = 39$, $n = 38$), *bri1* ($n = 28$, $n = 32$). Statistical significance was determined using linear mixed effect modelling followed by comparison of each genotype to the wild-type control using unpaired two-sided t-test followed by multiple testing correction using the Holm method. *apex-1* $^*P = 2.53 \times 10^{-14}$, *apex-2* $^*P = 1.10 \times 10^{-5}$, *apex-3* $^*P = 1.55 \times 10^{-12}$, *bri1* $^*P = 8 \times 10^{-16}$. **e**, flg22-induced oxidative bursts represented as total photon counts over 40 min. Genetic backgrounds are indicated. Dots represent individual observations from three independent experiments. n denotes numbers of biologically independent leaf discs: WT ($n = 31$), *apex-1* ($n = 19$), *apex-2* ($n = 23$), *apex-3* ($n = 25$), *fls2* ($n = 15$). Statistical significance was determined using linear mixed effect modelling followed by comparison of each genotype to the wild-type control using an unpaired two-sided t-test followed by multiple testing correction using the Holm method. *apex-1* $^*P = 2.99 \times 10^{-3}$, *apex-2* $^*P = 2.84 \times 10^{-2}$, *apex-3* $^*P = 2.84 \times 10^{-2}$, *fls2* $^*P = 8 \times 10^{-16}$. All box plots display the first and third quartiles, split by the median (red line); whiskers extend to include the maximum and minimum values.



Extended Data Figure 9 | Modulation of brassinosteroid signalling by AT5G51560. a, Morphology of representative seedlings grown for 7 days in the absence (NT) or presence (BL) of 500 nM brassinolide. Genotypes are indicated. The experiment was conducted twice with similar results. **b**, Hypocotyl length fold changes corresponding to **a**. Genotypes are indicated. Dots represent individual observations from two independent experiments. *n* denotes numbers of biologically independent hypocotyl: WT ($n = 39$ (NT), $n = 29$ (T)), *at5g51560* line 1 ($n = 36$ (NT), $n = 26$ (T)), *at5g51560* line 2 ($n = 39$ (NT), $n = 34$ (T)), *bri1* ($n = 25$ (NT), $n = 27$ (T)). Box plots display the first and third quartiles, split by the median; whiskers extend to include the maximum and minimum values. Statistical significance was determined using linear mixed effect modelling followed by comparison of each genotype to the wild-type control using an unpaired two-sided *t*-test followed by multiple testing correction using the Holm method. *at5g51560* line 1 $*P = 3.75 \times 10^{-6}$, *at5g51560* line 2 $*P = 2.26 \times 10^{-12}$, *bri1* $*P = 6 \times 10^{-16}$.

doi:10.1038/nature25187

High response rate to PD-1 blockade in desmoplastic melanomas

Zeynep Eroglu[1,2]*, Jesse M. Zaretsky[1]*, Siwen Hu-Lieskovan[1]*, Dae Won Kim[2,3], Alain Algazi[4], Douglas B. Johnson[5], Elizabeth Liniker[6], Ben Kong[7], Rodrigo Munhoz[8,9], Suthee Rapisuwon[10], Pier Federico Gherardini[11], Bartosz Chmielowski[1], Xiaoyan Wang[1], I. Peter Shintaku[1], Cody Wei[1], Jeffrey A. Sosman[5]†, Richard W. Joseph[12], Michael A. Postow[8,9], Matteo S. Carlino[6,7,13], Wen-Jen Hwu[3], Richard A. Scolyer[6,13,14], Jane Messina[2], Alistair J. Cochran[1], Georgina V. Long[6,13,15] & Antoni Ribas[1]

Desmoplastic melanoma is a rare subtype of melanoma characterized by dense fibrous stroma, resistance to chemotherapy and a lack of actionable driver mutations, and is highly associated with ultraviolet light-induced DNA damage[1]. We analysed sixty patients with advanced desmoplastic melanoma who had been treated with antibodies to block programmed cell death 1 (PD-1) or PD-1 ligand (PD-L1). Objective tumour responses were observed in forty-two of the sixty patients (70%; 95% confidence interval 57–81%), including nineteen patients (32%) with a complete response. Whole-exome sequencing revealed a high mutational load and frequent *NF1* mutations (fourteen out of seventeen cases) in these tumours. Immunohistochemistry analysis from nineteen desmoplastic melanomas and thirteen non-desmoplastic melanomas revealed a higher percentage of PD-L1-positive cells in the tumour parenchyma in desmoplastic melanomas ($P=0.04$); these cells were highly associated with increased CD8 density and PD-L1 expression in the tumour invasive margin. Therefore, patients with advanced desmoplastic melanoma derive substantial clinical benefit from PD-1 or PD-L1 immune checkpoint blockade therapy, even though desmoplastic melanoma is defined by its dense desmoplastic fibrous stroma. The benefit is likely to result from the high mutational burden and a frequent pre-existing adaptive immune response limited by PD-L1 expression.

Desmoplastic melanoma (DM) accounts for fewer than 4% of melanomas. It is characterized histologically by spindle-shaped melanoma cells within abundant collagenous stroma with scattered lymphoid aggregates, and typically has a high mutational burden from ultraviolet light radiation-induced damage[1]. Anti-PD-1 antibodies have been approved in many countries for the treatment of advanced melanoma, and have an overall response rate of 33–40%[2]. As recognition of neoantigens that result from somatic non-synonymous mutations is associated with improved clinical responses to anti-PD-1 and anti-PD-L1 therapy[3–6], we hypothesized that patients with DM might respond well to anti-PD-1 or anti-PD-L1 therapies, owing to their high mutational load.

We conducted a retrospective review of the pathology reports from 1,058 patients with advanced melanoma treated with anti-PD-1 or anti-PD-L1 immunotherapies between 2011 and 2016 at ten international sites with high-volume melanoma clinical trials. We identified 60 patients with advanced, unresectable DM who received PD-1 or PD-L1 blockade therapy (Extended Data Tables 1, 2). Thirty-five patients (58%) had extra-pulmonary visceral metastases or elevated lactate dehydrogenase (M1c disease), which are recognized makers of poor prognosis[7]. Local pathologists reported histological subclassification into pure ($n=25$), mixed ($n=30$) or indeterminate ($n=5$) DM subtypes[8]. All cases had the distinctive diagnostic features of DM with abundant connective tissue surrounding the tumour cells, which can be highlighted by Masson's trichrome stain (examples in Fig. 1a, with the collagenous stroma stained in blue). Central review of haematoxylin and eosin-stained tissue from 34 cases by two pathologists revealed that 65% of cases had a substantial fibrous stroma (graded 2–3), and that 63% of cases had lymphoid aggregates within the tumour and/or at the tumour stromal interface (graded 1–3) (Supplementary Table 1). Forty-two patients (70%) had progressed after prior systemic treatment, most frequently with the cytotoxic T lymphocyte antigen-4 (CTLA-4) blocking antibody ipilimumab (Extended Data Table 1 and Supplementary Table 1). The most frequently administered anti-PD-1 or anti-PD-L1 drug was pembrolizumab (in forty-five patients (75%)), while eight (13%) received nivolumab, three (5%) the anti-PD-L1 antibody BMS-936559, and an additional three (5%) received a combination of nivolumab or pembrolizumab with ipilimumab.

With a median follow up of 22 months, 42 out of the 60 patients (70%, 95% Clopper–Pearson confidence interval of 57–81%) had an objective response by RECIST 1.1 criteria (Fig. 1b, c). This included 19 (32%) complete responses and 23 (38%) partial responses; nine patients with a partial response eventually showed tumour progression but none of the patients with complete response did. When the four patients treated with a combination of anti-PD1 drugs and ipilimumab were excluded, responses were seen in 38 out of 56 (68%) patients. Three patients with isolated progression (including two who had a partial response) underwent surgery and subsequently had no evidence of melanoma with ongoing follow up for more than 1.8, 5.2, and 5.3 years. Median progression-free survival and overall survival have not been reached, with an estimated two-year overall survival of 74% (95% confidence interval 60–84%) (Extended Data Fig. 1a, b). For patients censored in the Kaplan–Meier curve, median follow-up was 27 months or more. There were no statistically significant differences in either objective response rate (65% versus 70%), or overall survival between patients with the two histological subtypes of DM (pure or mixed). There was also no difference in objective responses based on degree of fibrosis or presence of lymphoid aggregates (Supplementary Table 1).

Whole-exome sequencing from 17 cases in our DM cohort revealed more than 82% C>T transitions as part of a strong signature of ultraviolet light-induced DNA damage that is common to cutaneous melanoma[1,9] (Extended Data Fig. 2a, b). There was no difference in mutational load between locally advanced and metastatic lesions

[1]University of California Los Angeles, Los Angeles, California, USA. [2]Moffitt Cancer Center and University of South Florida, Tampa, Florida, USA. [3]The University of Texas-MD Anderson Cancer Center, Houston, Texas, USA. [4]University of California San Francisco, San Francisco, California, USA. [5]Vanderbilt Ingram Cancer Center, Nashville, Tennessee, USA. [6]Melanoma Institute Australia, Sydney, New South Wales, Australia. [7]Westmead Hospital, Sydney, New South Wales, Australia. [8]Memorial Sloan Kettering Cancer Center, New York, New York, USA. [9]Weill Cornell Medical College, New York, New York, USA. [10]Georgetown Lombardi Cancer Center, Washington DC, USA. [11]Parker Institute for Cancer Immunotherapy, San Francisco, California, USA. [12]Mayo Clinic, Jacksonville, Florida, USA. [13]The University of Sydney, Sydney, New South Wales, Australia. [14]Royal Prince Alfred Hospital, Sydney, New South Wales, Australia. [15]Royal North Shore Hospital, Sydney, New South Wales, Australia. †Present address: Northwestern University Medical Center, Chicago, Illinois, USA.
*These authors contributed equally to this work.

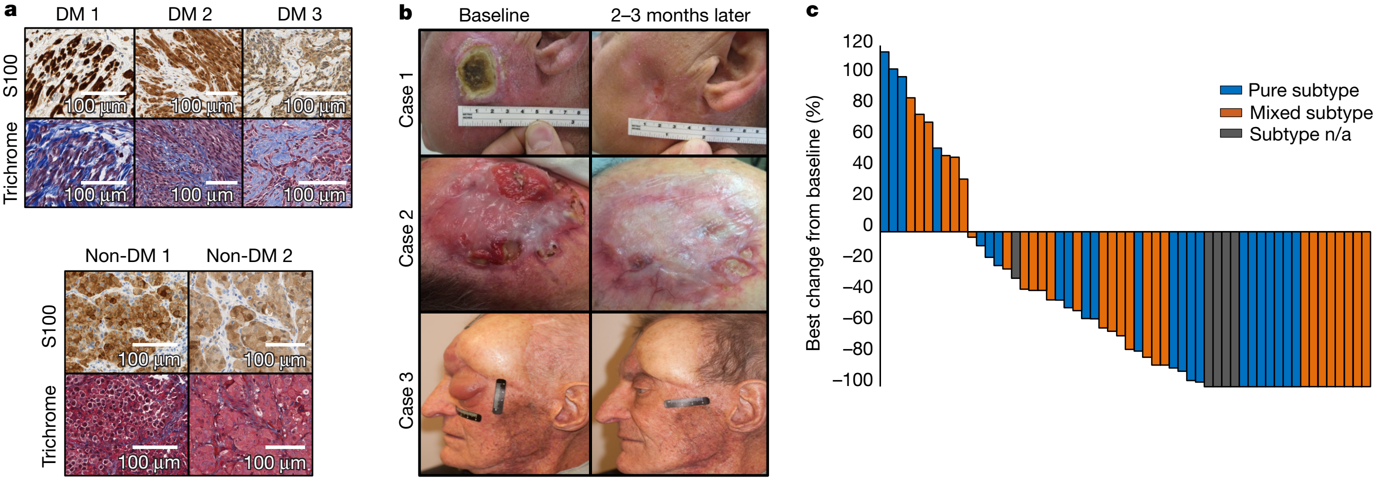


Figure 1 | High response rate to PD-1 blockade in patients with DM. **a**, Histological examples of three cases of DM (top) compared with two cases of non-desmoplastic cutaneous melanoma (non-DM, bottom) stained with Masson's trichrome stain (bottom rows) to highlight the collagenous stroma characteristic of DM. Top rows, S100 stains (brown). Bottom rows, Masson's trichrome stain (blue collagenous stroma, red cytoplasm and brown nuclei). **b**, Images of three cases of DM that responded to PD-1 blockade therapy. Left, baseline images (before treatment with anti-PD-1 or anti-PD-L1 therapy); right, images taken after 2–3 months of anti-PD-1 therapy. **c**, Waterfall plot of best response on therapy of 56 patients with DM treated with anti-PD-1 or anti-PD-L1 antibodies (data were not available (n/a) for four patients, three who had progressive disease and one who had a partial response). Images for case 1 in **b** reproduced with permission from[29] *New Engl. J. Med.*, Hamid, O. *et al.* Safety and tumour responses with lambrolizumab (anti-PD-1) in melanoma. **369**, 134–144 Copyright © 2013 Massachusetts Medical Society.

(Extended Data Fig. 3a). Mutations in *NF1* in the absence of *BRAF* or RAS family hotspot mutations were the most common driving genetic event (82.4%, 14 of 17 samples), along with enrichment for loss-of-function mutations in *TP53* and *ARID2* (Fig. 2a, Extended Data Fig. 3b), similar to previously published series of DM[1,10]. These features are also characteristic of *NF1* subtype melanoma, which comprises 8–12% of cutaneous melanoma cases in large cohorts and has more than double the mutational load of the *NRAS*, *BRAF* or triple wild-type subtypes[11–13]. Our DM series had similar mutational load to *NF1* subtype cases (regardless of histological classification) in a combined

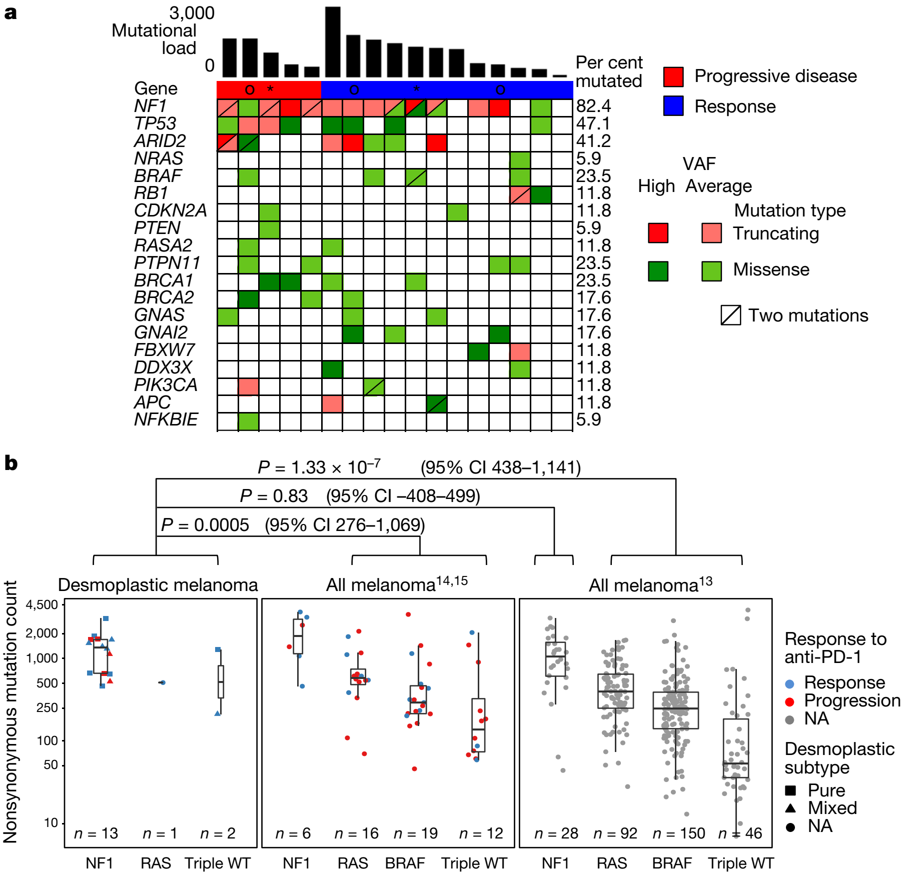


Figure 2 | High mutational load and similarity to *NF1* subtype in DM. **a**, Top bar graph represents mutational load. Tiling plot shows mutations in a given gene (rows) per sample (columns). In the tiling plot, top line represents response, as either primary resistance or progressive disease (red; $n = 5$), or response (partial or complete response and stable disease for more than 12 months; dark blue; $n = 12$). Colour indicates mutation type, with truncating mutations (frameshift, nonsense, splice-site) in red, missense in green. Darker colour intensity indicates potentially homozygous mutations, with variant allele frequency (VAF) more than 1.5 times the sample median. Asterisk, biopsy from responding lesion despite a mixed response and eventual progression. Circle, patient showed no evidence of disease for more than 1 year after surgical resection of a progressing lesion. **b**, Non-synonymous mutations determined by whole-exome sequencing from the current DM cohort, two pooled studies of anti-PD1 treated cutaneous melanoma[14,15] and TCGA data[13]. Each cohort is split by driver mutation subtype. Colour indicates PD1 blockade therapy response (red, progression; blue, response), and shape represents the subtype of DM (pure versus mixed). In the box plots, line shows median, box shows 25th and 75th percentiles, whiskers show highest and lowest values within 1.5 times interquartile range. Two-sided Wilcoxon Mann–Whitney rank sum test.

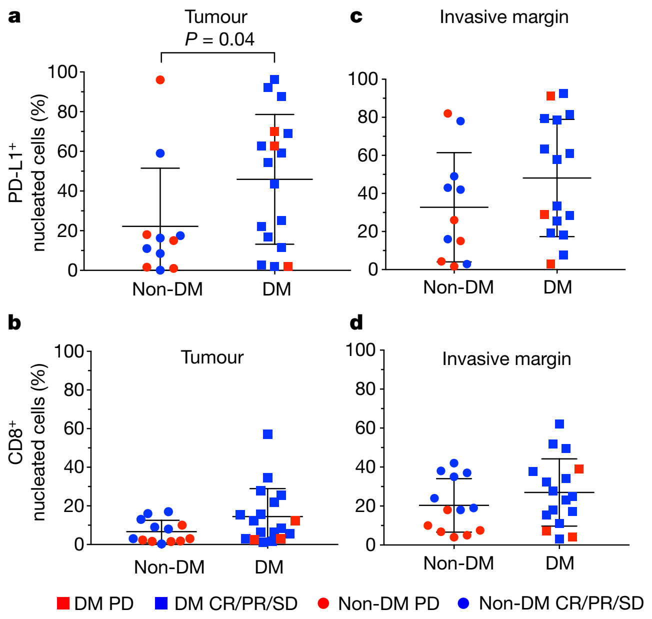


Figure 3 | CD8 density and PD-L1 expression in the tumour parenchyma and invasive margins from biopsies of patients with DM and non-DM tumours. a, PD-L1 staining in the tumour centre (non-DM: 1CR/5PR/5PD; DM: 7CR/6PR/1SD/3PD). **b**, CD8 staining in the tumour centre (non-DM: 2CR/5PR/6PD; DM: 7CR/7PR/1SD/3PD). **c**, PD-L1 staining in the invasive margin (non-DM: 1CR/5PR/5PD; DM: 6CR/6PR/1SD/3PD). **d**, CD8 staining in the invasive margin (non-DM: 2CR/5PR/6PD; DM: 6CR/7PR/1SD/3PD). Data show percentage of positively stained cells in all nucleated cells. PD, progressive disease; SD, stable disease; CR, complete response; PR, partial response. See Supplementary Table for all statistical analyses.

series from two reports of patients with anti-PD-1-treated advanced melanoma[14,15] and in data from the Cancer Genome Atlas (TCGA)[13]. In all three series, *NF1* mutated cases had a substantially greater mutational load than the non-*NF1* subtypes, but there was no difference in response to PD-1 blockade (Fig. 2b). Patients with DM that did not respond ($n = 5$) showed no difference in mutational load compared with patients that did respond (rank sum $P = 0.87$, Fig. 2b). This is consistent with the findings of two previous anti-PD-1-treated cohorts[14,15] but not with data from patients with melanoma treated with anti-CTLA4[16] (Extended Data Fig. 3c) or patients with lung and bladder cancer treated with anti-PD-1 or anti-PD-L1 therapy[3,6]. We did not find any genes that were mutated more frequently in patients with DM with or without response to therapy (Extended Data Fig. 4a), including when performing specific analyses for potential detrimental mutations in the interferon receptor pathway or *B2M* that may result in innate or acquired resistance to anti-PD-1 therapy[14,17] (Extended Data Fig. 4b).

We evaluated whether the presence of CD8$^+$ T cells and PD-L1 in DM was associated with response to anti-PD1 or anti-PD-L1 therapy[18,19] using 19 available pre-treatment DM tumour biopsies compared to 13 non-DM samples (seven with a complete or partial response, six with progressive disease) using digital quantitative immunohistochemistry (IHC). We used S100 expression to define the invasive tumour margin (stromal-tumour edge) and inside tumour parenchyma (tumour centre) (examples in Extended Data Fig. 5f.). Overall, biopsies from patients with DM had a notably higher percentage of PD-L1 positive cells in the tumour parenchyma than non-DM cases ($P = 0.04$, Fig. 3a), confirming the same finding from primary DM lesions[20]. There were no significant differences in the density of CD8$^+$ cells in the tumour parenchyma, or of CD8$^+$ and PD-L1$^+$ cells in the invasive margin ($P = 0.12$, $P = 0.41$ and $P = 0.16$, respectively; Fig. 3b–d). Consistent with previous observations[18], the strongest correlation with clinical benefit (defined as having a complete or partial response, or prolonged stable disease for more than 12 months) was baseline density of CD8$^+$ T cells in the invasive margin in non-DM melanoma ($P = 0.002$, Extended Data Fig. 6a–d).

In DM samples, PD-L1 expression in the tumour parenchyma was significantly associated with CD8 density ($P = 0.007$) and PD-L1 expression in the invasive margin ($P = 0.0003$), but not with CD8 density inside the tumour parenchyma ($P = 0.15$, Extended Data Fig. 7). Similarly, PD-L1 expression in the invasive margin was significantly associated with CD8 density in the invasive margin ($P = 0.0003$), CD8 density in the tumour parenchyma ($P = 0.04$), and PD-L1 expression in the tumour parenchyma ($P = 0.0003$). Among DM cases for which we had exome sequencing, we did not detect many of the genetic mechanisms reported to cause constitutive PD-L1 expression, including amplification of the PD-L1–PD-L2–JAK2 (PDJ) locus, mutations or amplification of *MYC* or *EGFR*, or disruption of *CDK5*[21–24]. The 3′ UTR of *CD274* (encoding PD-L1) was not well captured in our exome sequencing, and disruption could not be assessed[25]. Therefore, the higher PD-L1 expression in DM is likely to result from a reactive response to CD8$^+$ T cell infiltrates that reflect adaptive immune resistance[26].

We noted five distinct patterns of CD8$^+$ cell infiltration and PD-L1 expression in the invasive margin and tumour parenchyma; most patients who responded to therapy had one of the three patterns characterized by high CD8$^+$ T cells (twelve out of fourteen with DM and six out of seven with non-DM; Extended Data Fig. 5a–e). Patients without a tumour response tended to have low CD8$^+$ cells regardless of the status of PD-L1 (Extended Data Fig. 5g), although a small number of patients (two out of nine) whose tumours had low baseline CD8$^+$ infiltrates responded to therapy. We integrated the data regarding CD8 and PD-L1 expression in biopsies with response and mutational load, allowing cases of DM and non-DM to self-organize on the basis of these data (Extended Data Fig. 8a and b). CD8 and PD-L1 levels did not differ between cases with pure or mixed DM histology (Extended Data Fig. 8b). Biopsies in which the invasive margin showed higher CD8$^+$ density clustered together, usually with higher PD-L1 expression both in the tumour and in the invasive margin, and were enriched in patients with an objective tumour response. Mutational load, which was relatively high in all these cases, did not cluster with any particular pattern of CD8 or PD-L1 expression, or with response to therapy.

Dense collagenous stroma as found in DM has been thought to be an important limitation for immune infiltration, as has been described for pancreatic cancer[27]. However, our data challenge this notion, as there are pre-existing T cell infiltrates in the invasive edge of DM lesions, and DMs show a much higher response rate to anti-PD1 therapy than any other subtype of melanoma. The response rate of 70% in DM, together with relapsed Hodgkin's disease and Merkel cell carcinomas[21,28], is among the highest responses to single agent PD-1 blockade therapy in any pathologically defined cancer. Our data suggest that DM, and probably the non-DM *NF1* subtype arising from sun-exposed areas, have a high response rate to PD-1 blockade therapy because they have a more dynamic pre-existing adaptive immune response.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 February; accepted 30 November 2017.
Published online 10 January 2018.

1. Shain, A. H. *et al.* Exome sequencing of desmoplastic melanoma identifies recurrent NFKBIE promoter mutations and diverse activating mutations in the MAPK pathway. *Nat. Genet.* **47,** 1194–1199 (2015).
2. Ribas, A. *et al.* Association of pembrolizumab with tumor response and survival among patients with advanced melanoma. *J. Am. Med. Assoc.* **315,** 1600–1609 (2016).
3. Rizvi, N. A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348,** 124–128 (2015).
4. Le, D. T. *et al.* PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372,** 2509–2520 (2015).

5. Hugo, W. *et al.* Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* **165,** 35–44 (2016).
6. Rosenberg, J. E. *et al.* Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial. *Lancet* **387,** 1909–1920 (2016).
7. Han, D. *et al.* Clinicopathologic predictors of survival in patients with desmoplastic melanoma. *PLoS ONE* **10,** e0119716 (2015).
8. Busam, K. J. *et al.* Cutaneous desmoplastic melanoma: reappraisal of morphologic heterogeneity and prognostic factors. *Am. J. Surg. Pathol.* **28,** 1518–1525 (2004).
9. Alexandrov, L. B. Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
10. Wiesner, T. *et al.* NF1 mutations are common in desmoplastic melanoma. *Am. J. Surg. Pathol.* **39,** 1357–1362 (2015).
11. Krauthammer, M. *et al.* Exome sequencing identifies recurrent mutations in *NF1* and RASopathy genes in sun-exposed melanomas. *Nat. Genet.* **47,** 996–1002 (2015).
12. Hayward, N. K. *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature* **545,** 175–180 (2017).
13. Akbani, R. *et al.* Genomic classification of cutaneous melanoma. *Cell* **161,** 1681–1696 (2015).
14. Shin, D. S. *et al.* Primary resistance to PD-1 blockade mediated by JAK1/2 mutations. *Cancer Discov*. **7,** 188–201 (2017).
15. Roh, W. *et al.* Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci. Transl. Med*. **9,** eaah3560 (2017).
16. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350,** 207–211 (2015).
17. Zaretsky, J. M. *et al.* Mutations associated with acquired resistance to PD-1 blockade in melanoma. *N. Engl. J. Med.* **375,** 819–829 (2016).
18. Tumeh, P. C. *et al.* PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515,** 568–571 (2014).
19. Daud, A. I. *et al.* Programmed death-ligand 1 expression and response to the anti-programmed death 1 antibody pembrolizumab in melanoma. *J. Clin. Oncol.* **34,** 4102–4109 (2016).
20. Frydenlund, N. *et al.* Tumoral PD-L1 expression in desmoplastic melanoma is associated with depth of invasion, tumor-infiltrating CD8 cytotoxic lymphocytes and the mixed cytomorphological variant. *Mod. Pathol.* **30,** 357–369 (2017).
21. Ansell, S. M. *et al.* PD-1 blockade with nivolumab in relapsed or refractory Hodgkin's lymphoma. *N. Engl. J. Med.* **372,** 311–319 (2015).
22. Akbay, E. A. *et al.* Activation of the PD-1 pathway contributes to immune escape in EGFR-driven lung tumors. *Cancer Discov*. **3,** 1355–1363 (2013).
23. Casey, S. C. *et al.* MYC regulates the antitumor immune response through CD47 and PD-L1. *Science* **352,** 227–231 (2016).
24. Dorand, R. D. *et al.* Cdk5 disruption attenuates tumor PD-L1 expression and promotes antitumor immunity. *Science* **353,** 399–403 (2016).
25. Kataoka, K. *et al.* Aberrant PD-L1 expression through 3′-UTR disruption in multiple cancers. *Nature* **534,** 402–406 (2016).
26. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12,** 252–264 (2012).
27. Jiang, H. *et al.* Targeting focal adhesion kinase renders pancreatic cancers responsive to checkpoint immunotherapy. *Nat. Med.* **22,** 851–860 (2016).
28. Nghiem, P. T. *et al.* PD-1 blockade with pembrolizumab in advanced Merkel-cell carcinoma. *N. Engl. J. Med.* **374,** 2542–2552 (2016).
29. Hamid, O. *et al.* Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. *N. Engl. J. Med.* **369,** 134–144 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements This study was funded in part by the Grimaldi Family Fund, the Parker Institute for Cancer Immunotherapy, National Institutes of Health (NIH) grants R35 CA197633 and P01 CA168585, the Ressler Family Fund, the Samuels Family Fund and the Garcia-Corsini Family Fund (to A.R.). Z.E. was supported in part by the Moffitt Cancer Center NCI Skin SPORE (5P50CA168536) and Moffitt's Total Cancer Care Initiative and Collaborative Data Services (P30-CA076292) for this work. J.M.Z. is part of the UCLA Medical Scientist Training Program supported by NIH training grant GM08042. S.H.-L. was supported by a Young Investigator Award and a Career Development Award from the American Society of Clinical Oncology (ASCO), a Tower Cancer Research Foundation Grant, and a Dr. Charles Coltman Fellowship Award from the Hope Foundation. We acknowledge the Translational Pathology Core Laboratory (TPCL) and R. Guo, W. Li, J. Pang and M. H. Macabali from UCLA for blood and biopsy processing, and X. Li, L. Dong, J. Yoshizawa, and J. Zhou from the UCLA Clinical Microarray Core for sequencing expertise. G.V.L. is supported by an NHMRC Fellowship and The University of Sydney Medical Foundation. R.A.S. is supported by an NHMRC Fellowship.

Author Contributions Z.E., J.M.Z., S.H.-L. and A.R. developed the concepts. Z.E., S.H.-L, J.M.Z., and A.R. designed the experiments. Z.E., J.M.Z., S.H.-L. and A.R. interpreted the data. S.H.-L., I.P.S. and Z.E. performed IHC analyses. J.M.Z. performed genomic analyses. Z.E., A.R., B.C., D.W.K., A.A., D.B.J., E.L., B.K., R.M., S.R., J.A.S., R.J., M.A.P., M.S.C, W.-J.H., and G.V.L. clinically evaluated patients and contributed clinical data and tumour samples. R.A.S., J.M., and A.J.C. evaluated tumour samples. P.F.G. conducted the heat map analysis. X.W. performed statistical analyses. C.W. evaluated the non-DM clinical data. Z.E., J.M.Z., S.H.-L. and A.R. wrote the manuscript. S.H.-L. and A.R. supervised the project. All authors contributed to the manuscript and approved the final version.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to S.H.-L. (shu-lieskovan@mednet.ucla.edu) or A.R. (aribas@mednet.ucla.edu).

METHODS

Analysis of clinical data. To conduct this retrospective analysis, records of 1,058 patients with advanced melanoma treated with anti-PD-1 or anti-PD-L1 therapy were reviewed across ten institutions to identify those with a diagnosis of DM. Each institution conducted its own search to find patients who fit these criteria. The study was conducted under Institutional Review Board approval at each centre and complied with all relevant ethical regulations. All patients had signed a local written informed consent form for research analyses. Consent to obtain photographs was obtained. No statistical methods were used to predetermine sample size. The experiments were not randomized.

Immunohistochemistry (IHC) analyses. Patients were selected for IHC analysis if they had adequate pre-treatment tumour samples and had signed a local written informed consent form for research analyses. Tumour samples were obtained from eight institutions. Slides cut from frozen or FFPE tissue samples were stained with haematoxylin and eosin, Masson's trichrome stain, or anti-S100, anti-CD8, and anti-PD-L1 at the UCLA Anatomic Pathology Immunohistochemistry and Histology Laboratory (CLIA-certified). Antibodies used included rabbit polyclonal S100 (DAKO, 1:1,000 dilution, low pH retrieval), CD8 clone C8/144B (Dako, 1:100, low pH retrieval), and PD-L1 (Spring Biosciences, Sp142, 1:200, high pH retrieval). IHC was performed on Leica Bond III autostainer using Bond ancillary reagents and a Refine Polymer Detection system. Slides were examined for the presence of CD8 and PD-L1 within the tumour parenchyma and the connective tissue surrounding the tumour (invasive margin). We defined the invasive margin (or leading edge) as the interfaces between individual tumour bundles and the fibrotic regions, as opposed to the intra-tumour staining, which is within the capsule of individual tumours. All slides were scanned at an absolute magnification of ×200 (resolution of 0.5 μm per pixel). An algorithm was designed based on pattern recognition that quantified immune cells within S100-positive areas (tumour) and S100-negative areas (invasive margin). The algorithm calculated the percentage cellularity (% positive cells/all nucleated cells) using the Halo platform (Indica Labs). This analysis system was not able to differentiate between tumour cell or infiltrating immune cell PD-L1 staining[30]. Immunohistochemical variables were compared between biopsies of patients who responded or progressed on treatement using the Wilcoxon Mann–Whitney test.

Lymphocytic infiltrate and fibrosis analysis. We analysed available pathological samples from 34 cases to define their lymphoid inflammation and degree of fibrosis. There is no quantitative measure for these readouts, so we used a semi-quantitative pathological assessment. Examples of each grade were circulated to pathology reviewers to ensure reproducibility. The investigators were not blinded to allocation during experiments and outcome assessment. When available, metastatic lesions were graded by the same schema as primary samples, as not all patients had primary tumour samples available for quantification. The hallmark of lymphoid infiltration in DM is the presence of lymphoid nodules within and occasionally surrounding the tumour. Therefore, we developed the grading schema below to describe the location of these nodules within the tumours:

0: no lymphoid aggregates
1: lymphoid aggregates within tumour
2: lymphoid aggregates at tumour–stroma interface
3: lymphoid aggregates within tumour and at tumour–stroma interface

A grading schema was also developed to describe the degree of fibrosis in tumours:

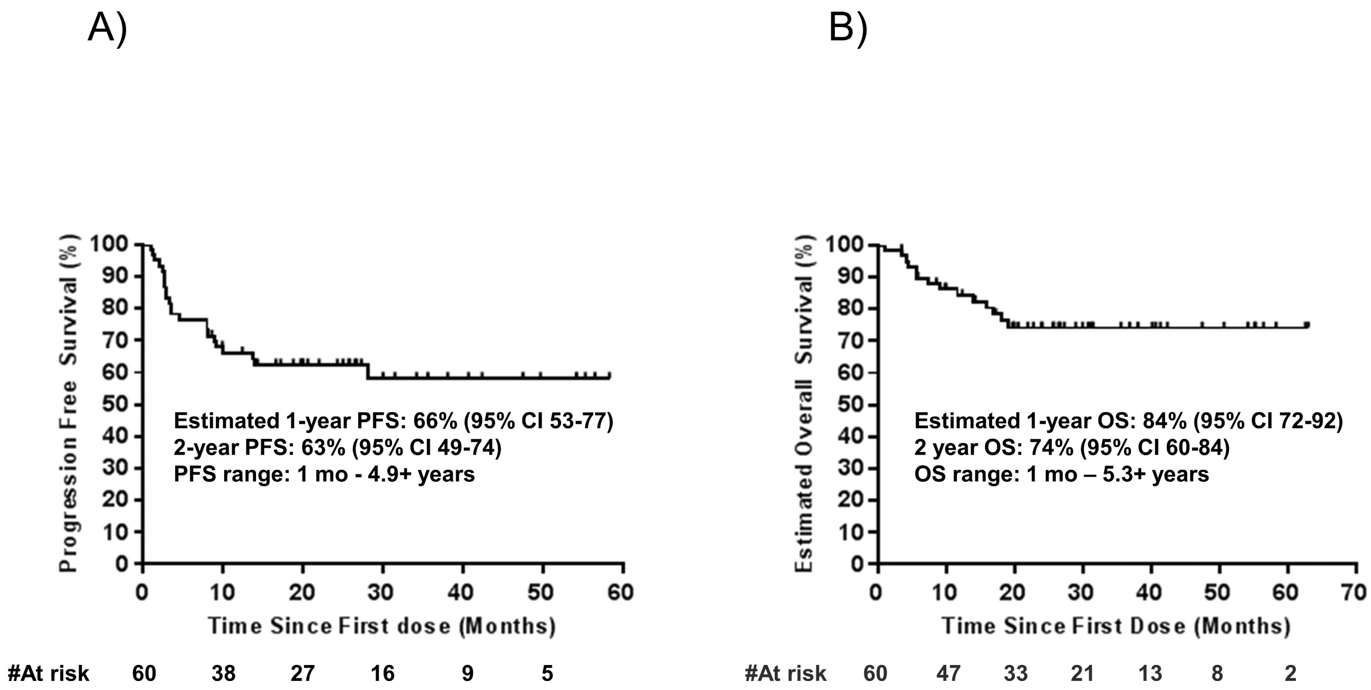
0: no significant stroma separates tumour cells
1: mild increase in fibroblasts and/or myxoid stroma separates tumour cells
2: moderate increase in fibroblasts and/or myxoid stroma separates tumour cells
3: tumour cells separated by abundant fibromyxoid stroma

Genetic analyses. In brief, whole-exome sequencing was performed at the UCLA Clinical Microarray Core using the Roche Nimblegen SeqCap EZ Human Exome Library v3.0 targeting 65 Mb of genome. Mutation calling was performed as previously described[14,17]. Out of 22 biopsies of DM sequenced, 17 cases (3 complete responses, 8 partial responses, 1 stable disease, 5 progressive disease) could be analysed by meeting quality control criteria for minimum coverage (50× tumour, 30× normal), tumour content (10%), and effective depth (coverage multiplied by tumour content >12×, representing >80% probability to detect heterozygous mutations with at least four reads). These were compared with exome sequencing from the TCGA[13], a prior DM cohort[1], and two anti-PD-1 monotherapy-treated cohorts, one from our group[14] with 23 cases which included a mix of responders and non-responders, and the second a subset of 30 patients after non-response to CTLA-4[15]. From that cohort to include one sample per patient, we excluded on-treatment samples in the setting of response; then we selected the biopsy with the highest tumour purity, regardless of time point, since most patients with more than one biopsy had <10% variance in their mutational loads. Response was defined as CR, PR, or SD for >12 months by RECIST1.1 in both cohorts. Mutation calling methods between cohorts all used MuTect at their core, and only non-synonymous mutations (Nonsense, Missense, Splice_Site, Frameshift indels, In-frame indels, Start_Codon indels or SNPs, and Stoploss/Nonstop variants) were assessed to minimize differences between exon-capture kits. An additional filter was applied to all data sets to exclude mutations at sites of known germline variation with an allele frequency >0.0005 in the Exome Aggregation Consortium (ExAC) database v0.3.1. Tumour purity was estimated by Sequenza, or as 2 × median variant allele frequency if less than 30%. Loss-of-function burden was determined using the LOF SIgRank algorithm[1], with the simulation run for 1,000 iterations and synonymous mutations for background mutation rate defined as silent, 3′UTR, 5′UTR, or exon-flanking intronic mutations. Single nucleotide variants and their flanking contexts were analysed for mutation signatures for the DM and UCLA non-DM[14] cohorts together using a published tool[9].

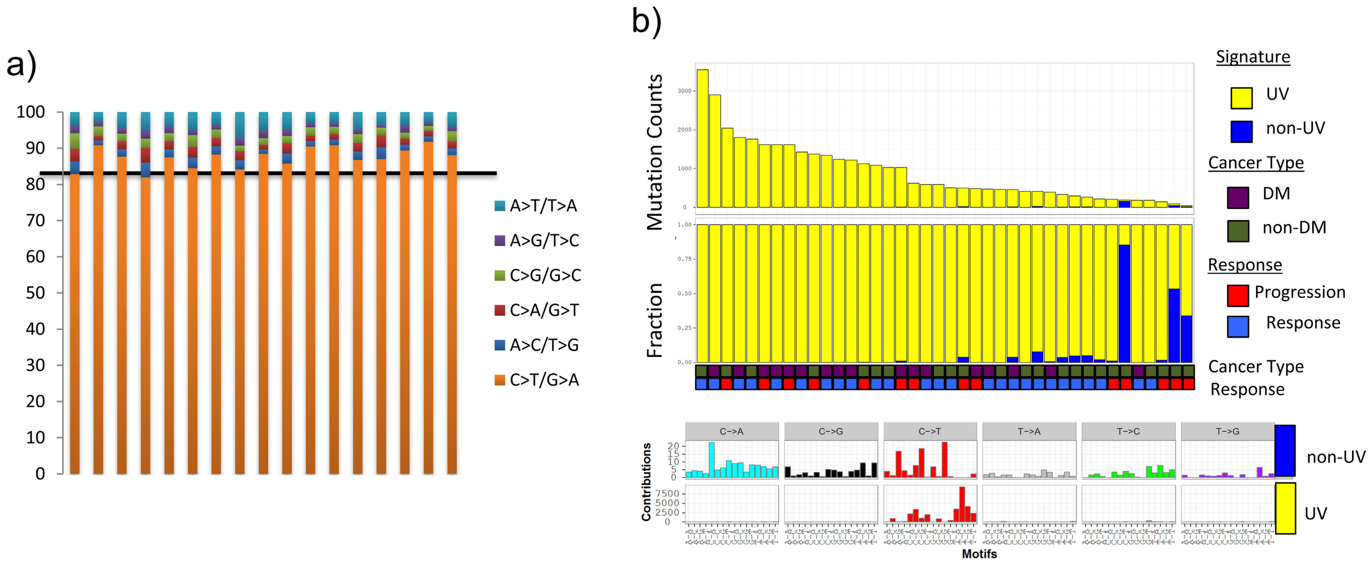
Statistical analyses. The Kaplan–Meier method and Greenwood's formula were used to estimate survival probabilities (survival rates and overall survival) and the corresponding 95% confidence intervals (CIs). Progression-free survival was defined from start of treatment to disease progression or death from any cause. Overall survival was defined from start of treatment to death from any cause. The objective response rate was reported as proportion along with Clopper–Pearson exact CIs. The chi-square and Fisher's exact tests were used to test for differences between groups for categorical variables. The Wilcoxon Mann–Whitney rank sum test was used to compare mutation rates between groups. Statistical analyses of the pathological data were performed using GraphPad Prism and mutation data using R v3.2.5. All tests were two-sided; *P* values <0.05 were considered statistically significant.

Data availability. Whole-exome sequencing data has been deposited in the National Center for Biotechnology Information (NCBI) dbGaP (https://www.ncbi.nlm.nih.gov/gap) with accession number phs001469. All other data are available from the authors on reasonable request.

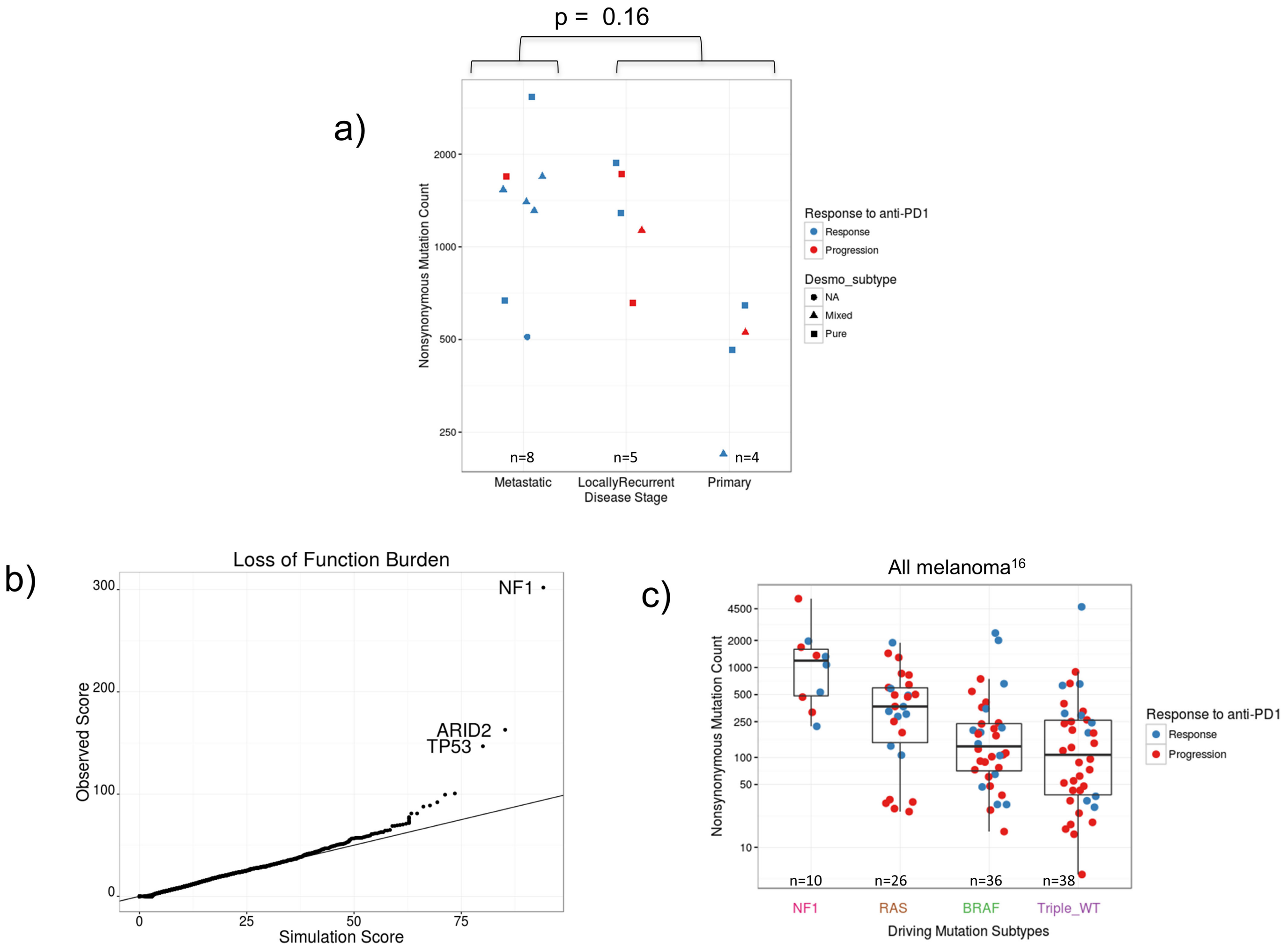
30. Rittmeyer, A. *et al.* Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet* **389,** 255–265 (2017).



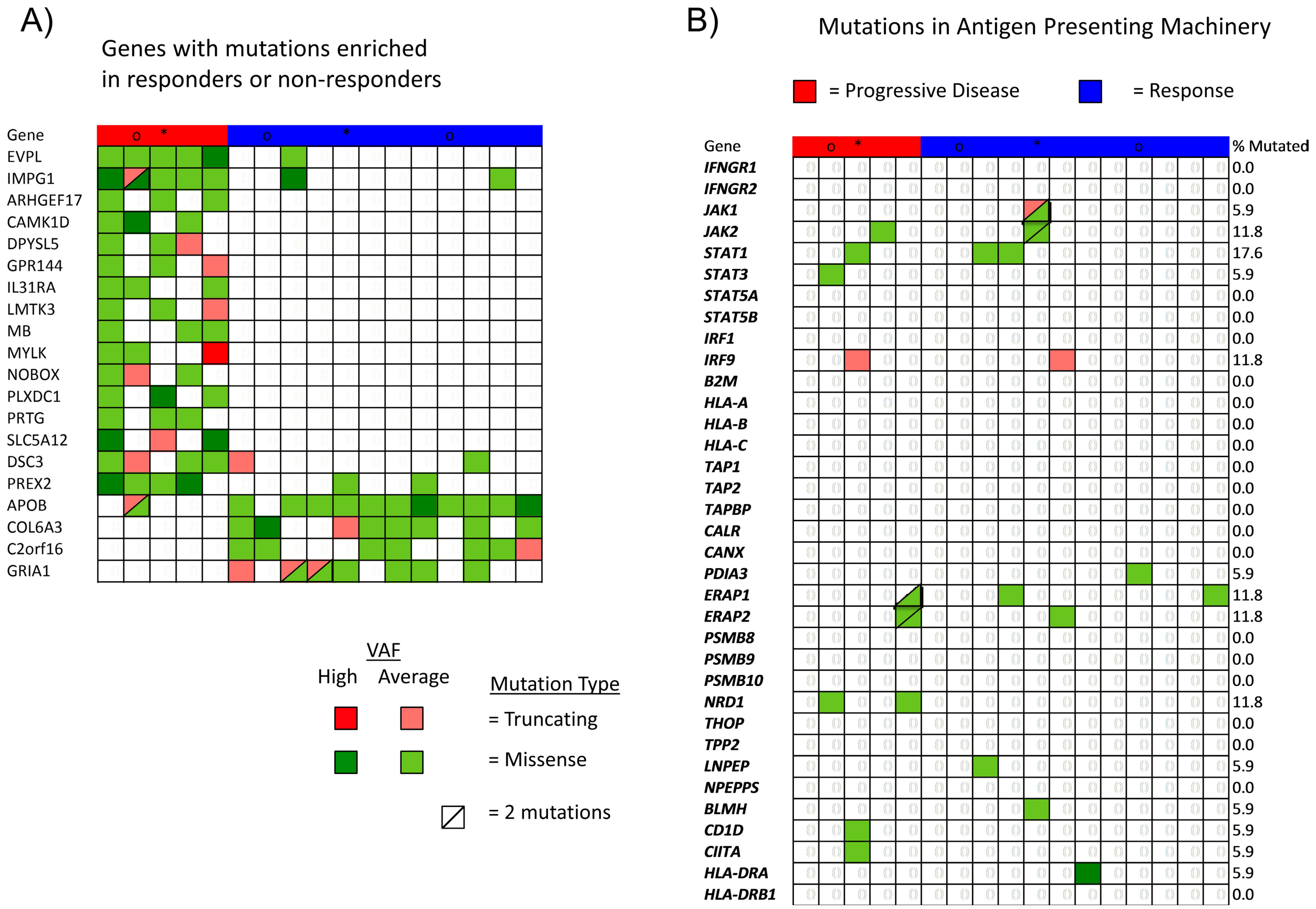
Extended Data Figure 1 | Survival data for the DM cohort. a, Progression-free survival (PFS), *n* = 60, median not reached. **b**, Overall survival (OS), *n* = 60, median not reached.



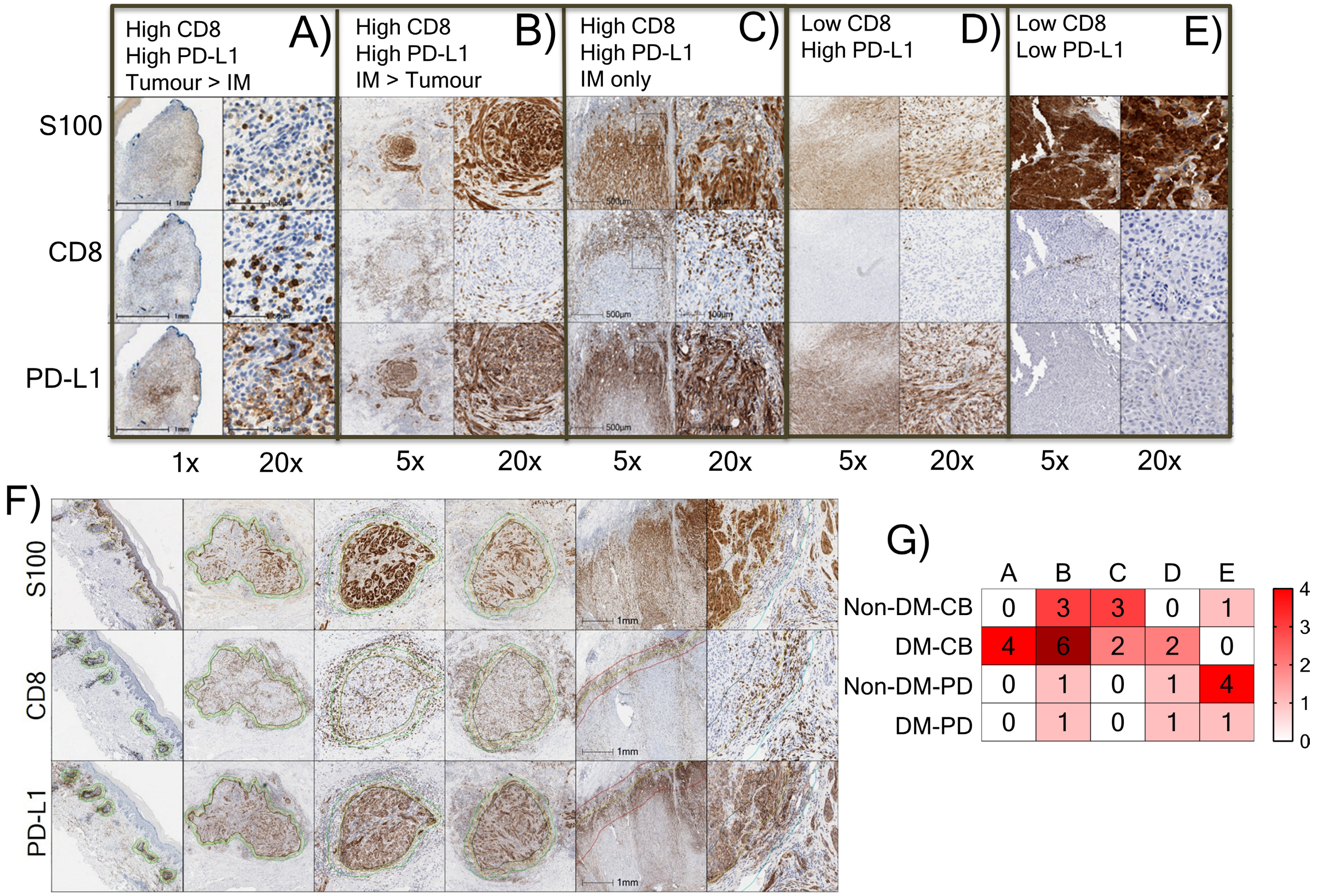
Extended Data Figure 2 | Ultraviolet light-induced DNA damage signature in the desmoplastic melanoma cohort. a, Cumulative percentage per DM sample ($n = 17$) of single nucleotide mutations by transition or transversion substitution. **b**, Mutation signature analysis[9] on combined DM ($n = 17$) and non-DM ($n = 23$) cohorts[14]. All show the predominant C>T-rich signature characteristic of UV damage.



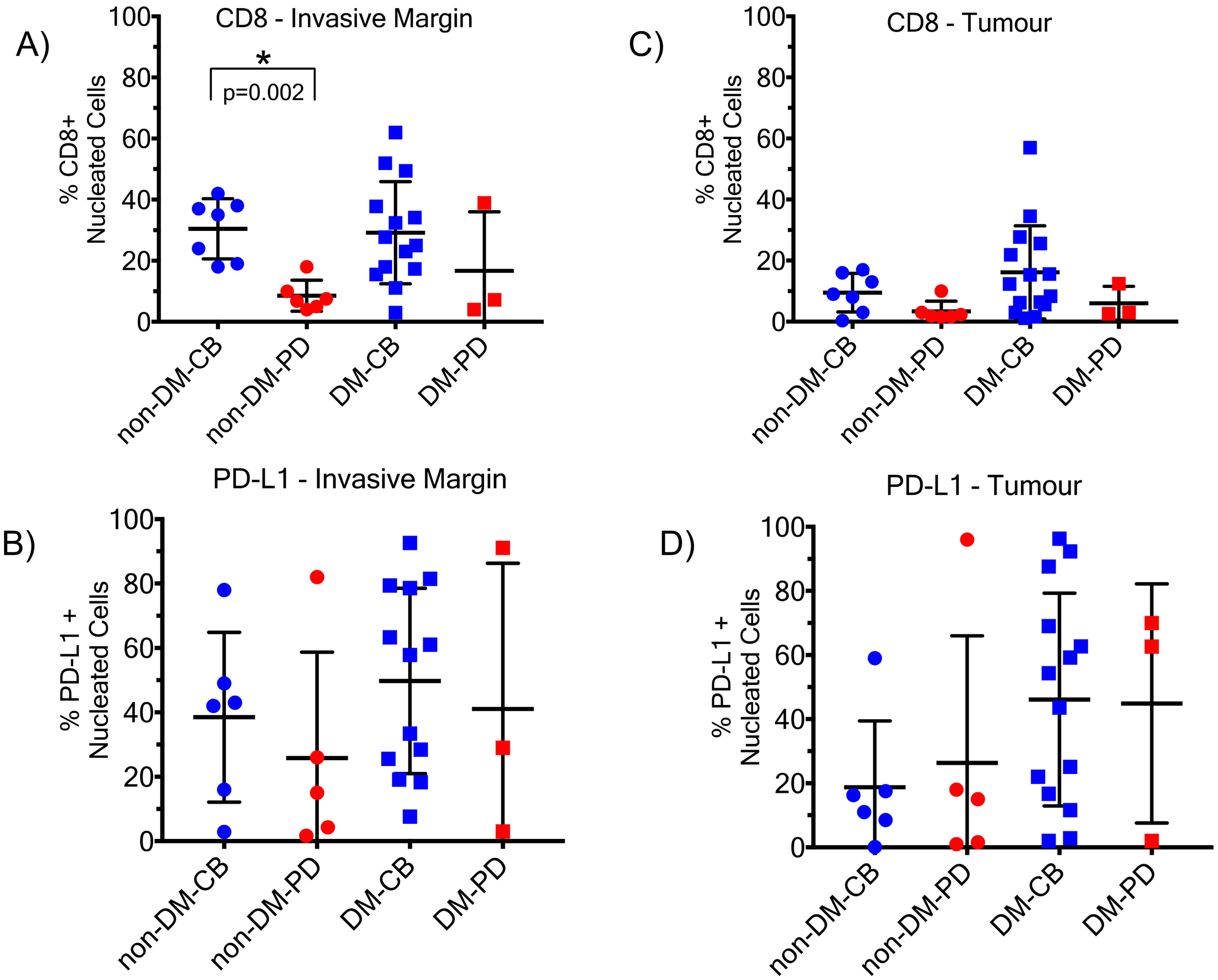
Extended Data Figure 3 | Mutational analysis in the desmoplastic melanoma cohort. a, Analysis of mutational load in samples obtained from primary locally advanced cases and metastatic lesions. Two-sided Wilcoxon Mann–Whitney rank sum test, *P* = 0.16 (95% CI, −171 to 1,175). **b**, Scores from the loss-of-function (LOF) SigRank algorithm[1] show enrichment for LOF mutations (nonsense, frameshift, splice-site or damaging missense) compared to the expected number based on the rate of LOF mutations in the cohort. Solid line corresponds to observed/expected ratio of 1.0. **c**, Mutational load in the vanAllen[16] anti-CTLA4 treated cohort separated by driver subtype and coloured by response. In the box plots, line is median, box is 25th to 75th percentile, whiskers show highest and lowest values within 1.5 × interquartile range.



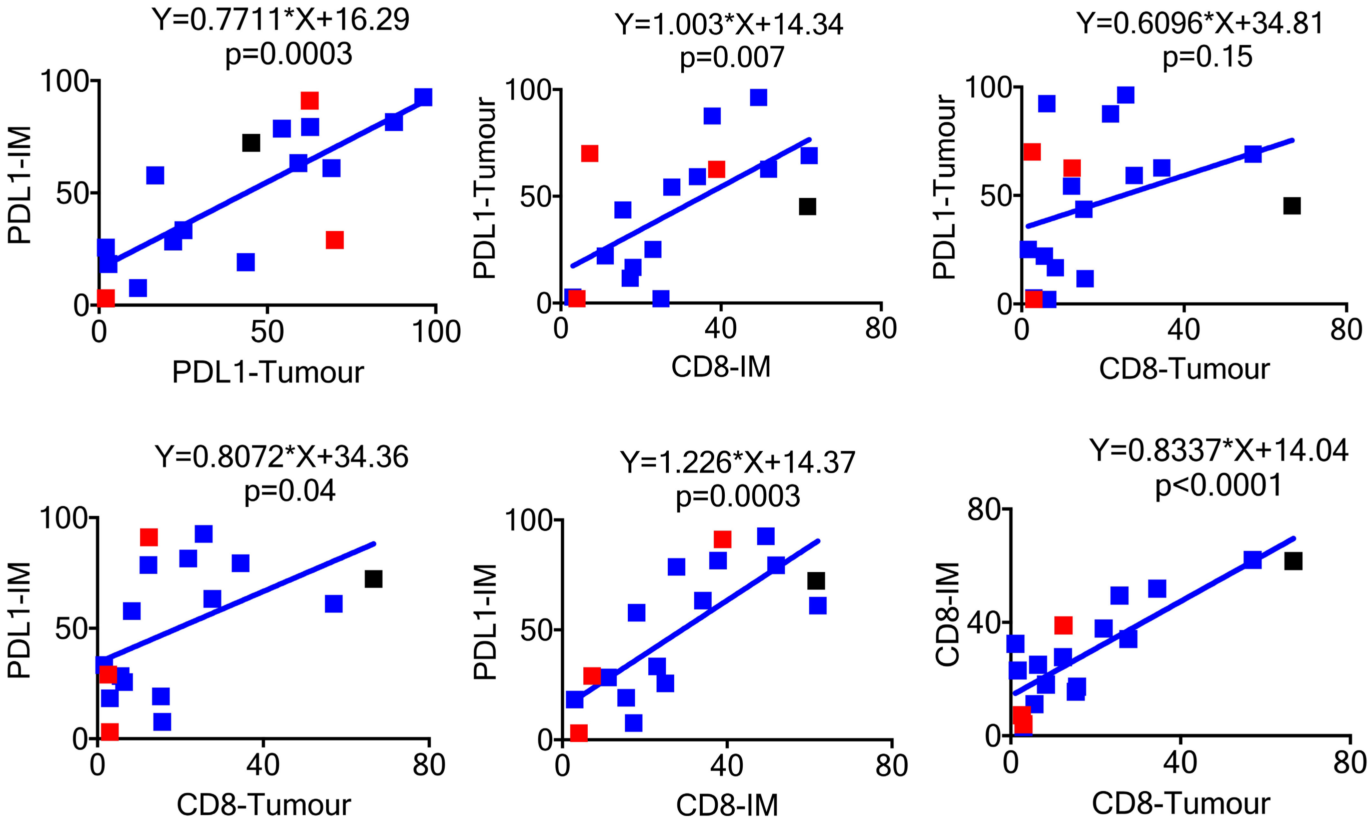
Extended Data Figure 4 | Mutations in antigen-presenting machinery or enriched by response in the DM cohort. a, Mutations in genes enriched in responders ($n = 12$) (blue) or non-responders ($n = 5$) (red). Shown are genes with $P < 0.05$ by unadjusted two-sided Fisher's exact test of samples with or without a non-synonymous mutation between responders and non-responders. None were significant after false-discovery rate adjustment. **b**, Mutations in antigen-presenting machinery genes. Tiling plot shows mutations in a given gene (rows) per sample (columns). Colour indicates mutation type, with truncating mutations (frameshift, nonsense, splice-site) in red, missense in green. Darker colour intensity indicates potentially homozygous mutations, with variant allele frequency more than 1.5 times the sample median.



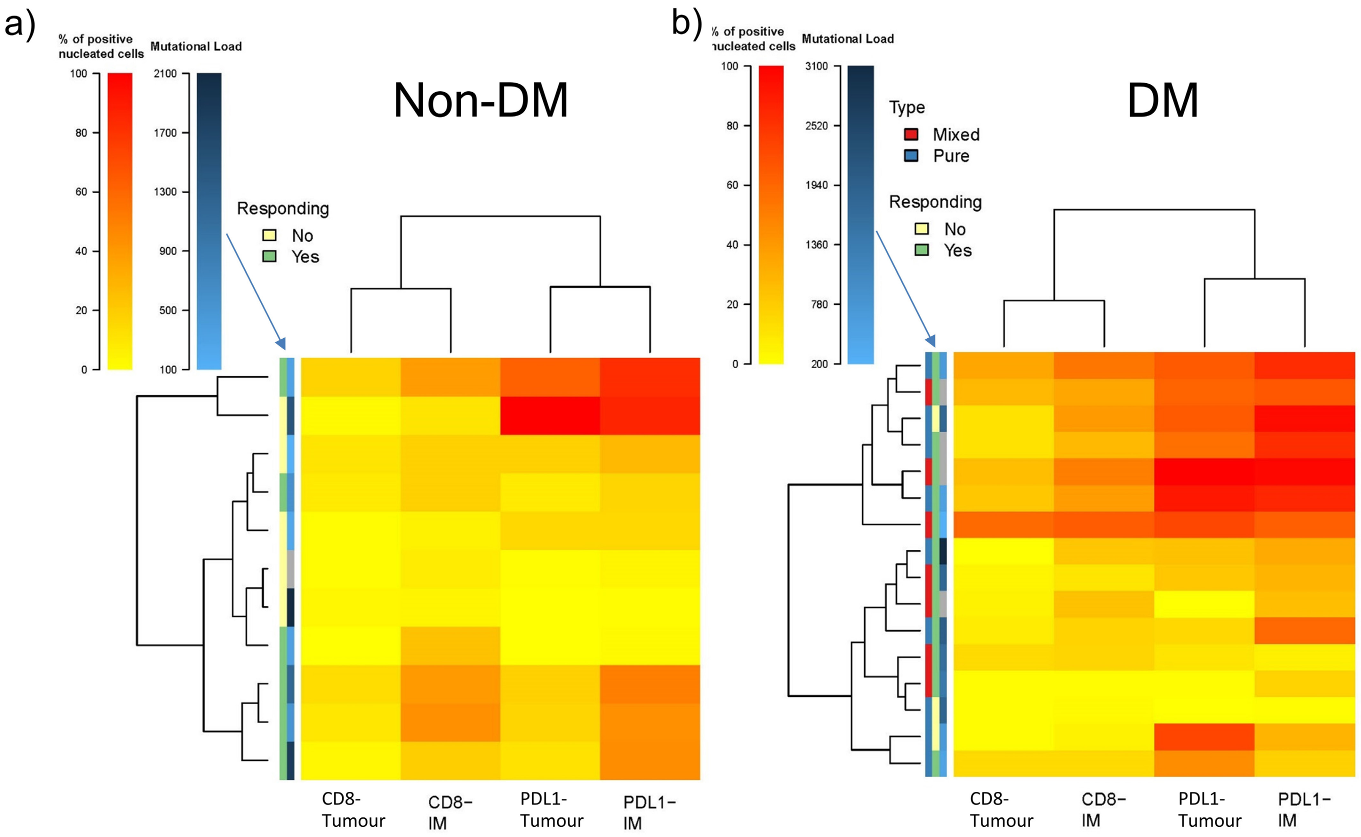
Extended Data Figure 5 | Patterns of CD8 infiltration and PD-L1 expression in biopsies from patients with DM and non-DM tumours. **a**–**e**, Using cut off of >10% for high CD8 density in either parenchyma or invasive margins and >15% for high PD-L1 expression, five different patterns were identified. **a**, High CD8 density, high PD-L1 in tumour parenchyma higher than in invasive margins. **b**, High CD8 density, high PD-L1 in invasive margins higher than in tumour parenchyma. **c**, High CD8 density, high PD-L1 in the invasive margins only. **d**, Low CD8 density, high PD-L1. **e**, Low CD8 density, low PD-L1 expression. **f**, Yellow lines delineate the edges of tumour regions determined by positive S100 staining. Green or red lines mark the invasive margins around the tumour edges. All analysis was done with HALO software (Indica Labs). **g**, Heat map summary of patterns of CD8 and PD-L1 expression in biopsies from patients with DM and CM, based on their response to anti-PD-1 or anti-PD-L1 treatment. Intensity of colour coding indicates number of cases in each category. All calculations were based on scanned whole tumour images.



Extended Data Figure 6 | CD8 density and PD-L1 expression in the tumour parenchyma and invasive margins in biopsies of patients with DM and non-DM tumours. a, CD8 staining in the invasive margin. **b**, PD-L1 staining in the invasive margin. **c**, CD8 staining in the tumour centre. **d**, PD-L1 staining in the tumour centre. The percentage of positively stained cells in all nucleated cells is shown. CB, clinical benefit; PD, progressive disease. All calculations used two-sided Mann–Whitney rank sum test. See Supplementary Table for all statistical analyses. Asterisk indicates statistical significance. Tumour, tumour centre.

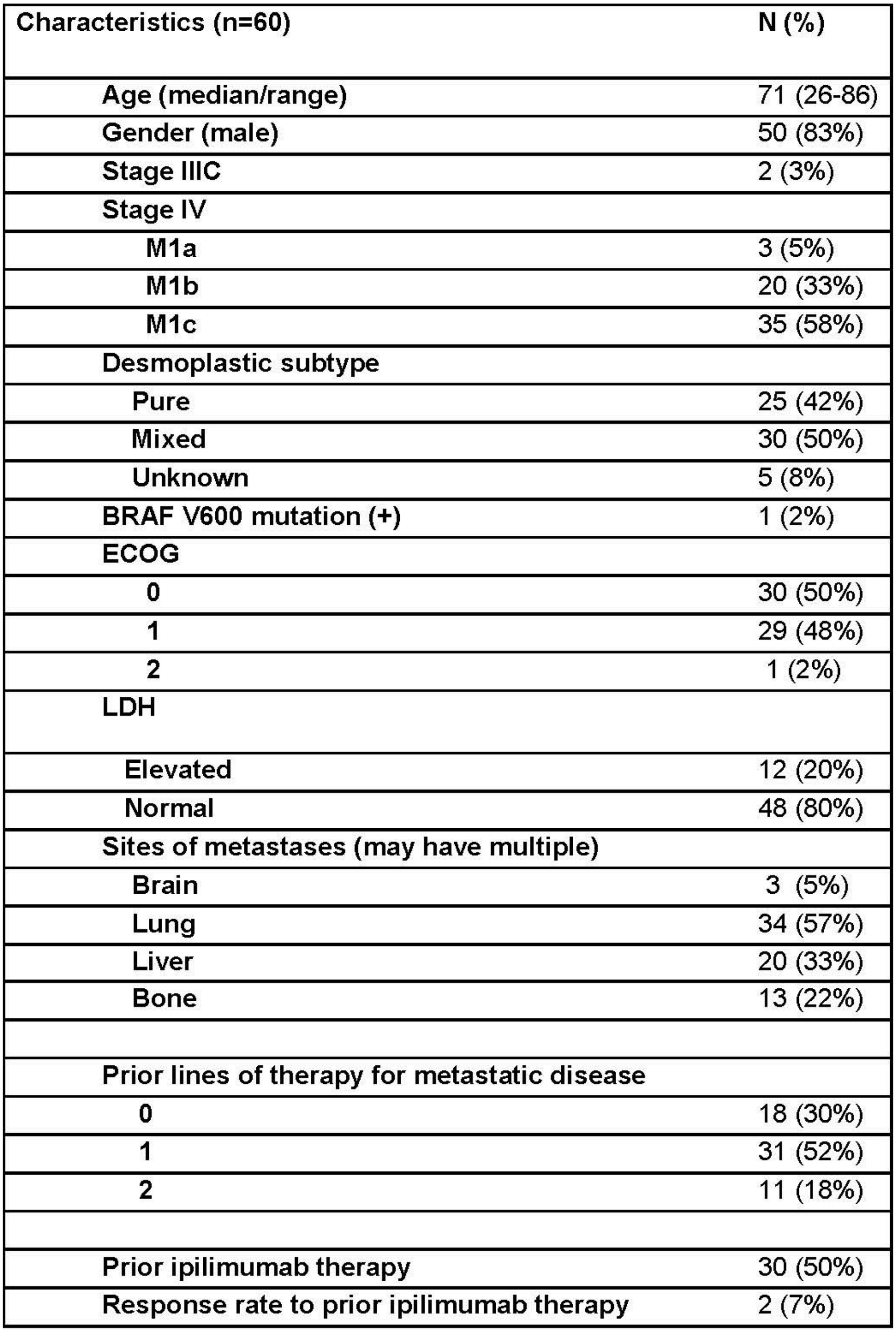


Extended Data Figure 7 | Correlation between CD8 and PD-L1 in the invasive margin or tumour parenchyma in DM. Black squares represent a sample from a patient who had a good response in the lesion biopsied (analysed) but was found to have brain metastasis shortly after treatment started. See Supplementary Table for further statistical analyses. IM, invasive margin.



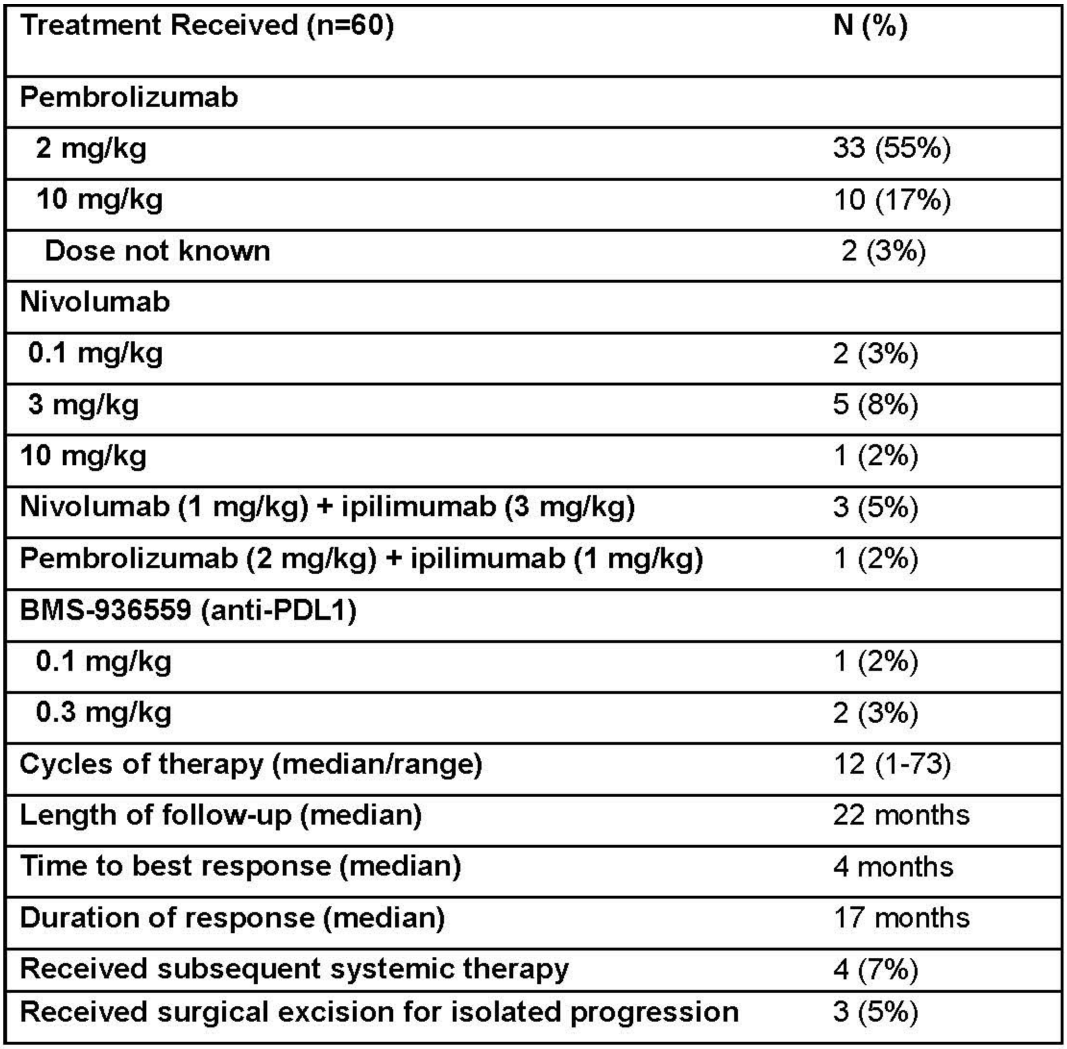
Extended Data Figure 8 | Hierarchical clustering of cases of DM and non-DM based on CD8 and PD-L1 expression in the invasive margin and tumour parenchyma. a, Non-desmoplastic cutaneous melanomas ($n = 13$), with the *y* axis colour coded for response and mutational load. **b**, Desmoplastic melanomas ($n = 19$), with the additional information of differentiation between pure (red) and mixed (blue) histology on the *y* axis. For mutational load, darker squares correspond to higher mutational load. Gray squares are missing data points.

Extended Data Table 1 | Summary of patient characteristics



Characteristics (n=60)	N (%)
Age (median/range)	71 (26-86)
Gender (male)	50 (83%)
Stage IIIC	2 (3%)
Stage IV	
M1a	3 (5%)
M1b	20 (33%)
M1c	35 (58%)
Desmoplastic subtype	
Pure	25 (42%)
Mixed	30 (50%)
Unknown	5 (8%)
BRAF V600 mutation (+)	1 (2%)
ECOG	
0	30 (50%)
1	29 (48%)
2	1 (2%)
LDH	
Elevated	12 (20%)
Normal	48 (80%)
Sites of metastases (may have multiple)	
Brain	3 (5%)
Lung	34 (57%)
Liver	20 (33%)
Bone	13 (22%)
Prior lines of therapy for metastatic disease	
0	18 (30%)
1	31 (52%)
2	11 (18%)
Prior ipilimumab therapy	30 (50%)
Response rate to prior ipilimumab therapy	2 (7%)

Extended Data Table 2 | Summary of systemic drug treatments received by each patient



Treatment Received (n=60)	N (%)
Pembrolizumab	
2 mg/kg	33 (55%)
10 mg/kg	10 (17%)
Dose not known	2 (3%)
Nivolumab	
0.1 mg/kg	2 (3%)
3 mg/kg	5 (8%)
10 mg/kg	1 (2%)
Nivolumab (1 mg/kg) + ipilimumab (3 mg/kg)	3 (5%)
Pembrolizumab (2 mg/kg) + ipilimumab (1 mg/kg)	1 (2%)
BMS-936559 (anti-PDL1)	
0.1 mg/kg	1 (2%)
0.3 mg/kg	2 (3%)
Cycles of therapy (median/range)	12 (1-73)
Length of follow-up (median)	22 months
Time to best response (median)	4 months
Duration of response (median)	17 months
Received subsequent systemic therapy	4 (7%)
Received surgical excision for isolated progression	3 (5%)

doi:10.1038/nature25170

Pharmacological activation of REV-ERBs is lethal in cancer and oncogene-induced senescence

Gabriele Sulli[1], Amy Rommel[2], Xiaojie Wang[3], Matthew J. Kolar[4], Francesca Puca[5], Alan Saghatelian[4], Maksim V. Plikus[3], Inder M. Verma[2] & Satchidananda Panda[1]

The circadian clock imposes daily rhythms in cell proliferation, metabolism, inflammation and DNA damage response[1,2]. Perturbations of these processes are hallmarks of cancer[3] and chronic circadian rhythm disruption predisposes individuals to tumour development[1,4]. This raises the hypothesis that pharmacological modulation of the circadian machinery may be an effective therapeutic strategy for combating cancer. REV-ERBs, the nuclear hormone receptors REV-ERBα (also known as NR1D1) and REV-ERBβ (also known as NR1D2), are essential components of the circadian clock[5,6]. Here we show that two agonists of REV-ERBs—SR9009 and SR9011—are specifically lethal to cancer cells and oncogene-induced senescent cells, including melanocytic naevi, and have no effect on the viability of normal cells or tissues. The anticancer activity of SR9009 and SR9011 affects a number of oncogenic drivers (such as HRAS, BRAF, PIK3CA and others) and persists in the absence of p53 and under hypoxic conditions. The regulation of autophagy and *de novo* lipogenesis by SR9009 and SR9011 has a critical role in evoking an apoptotic response in malignant cells. Notably, the selective anticancer properties of these REV-ERB agonists impair glioblastoma growth *in vivo* and improve survival without causing overt toxicity in mice. These results indicate that pharmacological modulation of circadian regulators is an effective antitumour strategy, identifying a class of anticancer agents with a wide therapeutic window. We propose that REV-ERB agonists are inhibitors of autophagy and *de novo* lipogenesis, with selective activity towards malignant and benign neoplasms.

The cell-autonomous circadian clock pleiotropically coordinates a complex network of physiological processes[1]. In both mice and humans, disruption of circadian rhythms increases cancer incidence[1,7]. Given the unique ability of the circadian clock to directly control several pathways that are crucial for tumorigenesis[2,8–11], pharmacological modulation of circadian components may offer promising selective anticancer strategies.

REV-ERBs are haem-binding circadian clock components[6,12,13] that act as repressors of processes involved in tumorigenesis, including metabolism[5,14,15], proliferation[16] and inflammation[2]. Binding to tetrapyrrole haem enhances the repressive function of REV-ERBs[13]. The development of the pyrrole derivatives SR9009 and SR9011[14] as specific agonists of REV-ERBs, with potent *in vivo* activity, prompted us to investigate whether pharmacological activation of these circadian repressors affects cancer cell viability by restraining pathways that are aberrantly activated in cancer.

SR9009 had a cytotoxic effect on cancer cells derived from a range of tumour types, namely brain cancer, leukaemia, breast cancer, colon cancer and melanoma (Fig. 1a, d, g, j, o). SR9011 displayed similar cytotoxic properties against the same cancer cell lines (Extended Data Fig. 1a–j). Notably, SR9009 and SR9011 are effective against tumour cell lines that harbour a range of oncogenic drivers, including HRAS, KRAS, BRAF, PTEN deficiency and β-catenin (Fig. 1, Extended Data Fig. 1), but have little or no toxic effect on normal cells at comparable concentrations (Fig. 1a, b, Extended Data Fig. 1a, b). Therefore, the antitumour activity of REV-ERB agonists is not limited to a single oncogenic driver, but is instead effective against a broad spectrum of tumorigenic pathways.

Levels of REV-ERB mRNA are comparable between normal cells and their transformed counterparts (Fig. 1c). The anticancer activity of SR9009 and SR9011 is abolished following the downregulation of REV-ERBs by multiple short hairpin RNAs (shRNAs) (Fig. 1o, p, Extended Data Fig. 1d, l).

The impairment of cancer cell viability on treatment with SR9009 and SR9011 is due to induction of apoptosis, as assessed by cleaved caspase 3 and terminal deoxynucleotidyl transferase (TdT) dUTP nick-end labelling (TUNEL) assays and further verified by electron microscopy (Fig. 1e, f, h, i, k, l, Extended Data Fig. 1g–k). As the tumour suppressor p53 has an important role in apoptosis and is often inactivated in cancer, we tested whether the induction of apoptosis by agonists of REV-ERBs requires p53. Agonist-induced apoptosis was largely intact in cells with compromised p53 function (mutation, deletion or shRNA-mediated downregulation; Fig. 1a, b, Extended Data Fig. 2a–j), which indicates that the downstream apoptosis trigger is independent of p53. Agonists of REV-ERBs do not, therefore, require the presence of wild-type p53 and are effective against several oncogenic pathways; these observations expand the potential therapeutic repertoire of agonists of REV-ERBs against multiple tumour types.

The selectivity of agonists of REV-ERBs towards cancer cells suggests that SR9009 and SR9011 may affect cellular processes that are critical specifically for the survival of tumour cells, and not essential for normal cells. The increased production of reactive oxygen species (ROS) is detrimental specifically to cancer cells[17], insofar as normal cells exhibit a greater tolerance for increased ROS production than do cancer cells. Agonists of REV-ERBs and other circadian clock components regulate mitochondrial metabolism and its oxidative activity[15,18]. If ROS overproduction is involved in the enhanced sensitivity of cancer cells to agonists of REV-ERBs, lowering oxidative stress would protect them against the agonists. We co-treated cancer cells with agonists of REV-ERBs and the antioxidant *N*-acetyl-L-cysteine (NAC). As a second way of lowering oxidative stress, we administered agonists of REV-ERBs under hypoxic conditions. In neither experimental setting was the ability of agonists of REV-ERBs to trigger apoptosis in cancer cells impaired (Extended Data Figs 2k–n, 3), which suggests that excessive ROS production is not involved in the enhanced sensitivity of cancer cells to these agonists.

Next we investigated whether agonists of REV-ERBs target anabolic pathways that are selectively critical for cancer cell survival. REV-ERBs tightly control lipid metabolism by repressing several lipogenic enzymes, including fatty acid synthase (FAS) and stearoyl-CoA

[1]Regulatory Biology Laboratory, Salk Institute for Biological Studies, La Jolla, California 92037, USA. [2]Laboratory of Genetics, Salk Institute for Biological Studies, La Jolla, California 92037, USA. [3]Department of Developmental and Cell Biology, University of California, Irvine, Irvine, California 92697, USA. [4]Clayton Foundation Laboratories of Peptide Biology, Salk Institute for Biological Studies, La Jolla, California 92037, USA. [5]Department of Genomic Medicine, The University of Texas MD, Anderson Cancer Center, Houston, Texas 77030, USA.

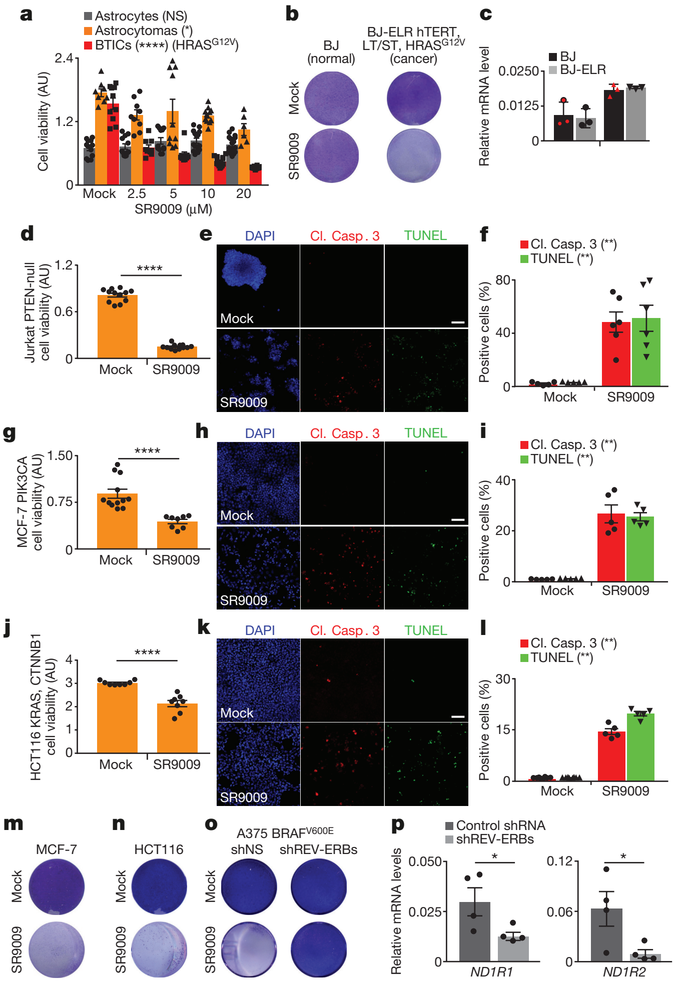


Figure 1 | SR9009 is selectively lethal in cancer cell lines driven by different oncogenic signalling. a, SR9009 treatment is cytotoxic specifically in cancer cells (72 h). One-way ANOVA. *n* indicates biological replicates: astrocytes, *n* = 12 (mock), *n* = 12 (2.5 µM), *n* = 12 (5 µM), *n* = 15 (10 µM), *n* = 18 (20 µM); astrocytomas, *n* = 8 (mock), *n* = 9 (2.5 µM), *n* = 10 (5 µM), *n* = 11 (10 µM) and *n* = 6 (20 µM), **P* = 0.037; and brain-tumour-initiating cells (BTICs), *n* = 10 (mock), *n* = 9 (2.5 µM), *n* = 9 (5 µM), *n* = 15 (10 µM) and *n* = 18 (20 µM), *****P* < 0.0001. **b**, SR9009 treatment impairs the viability of BJ-ELR, but not BJ, cells (proliferation assay, 7 days, 20 µM). LT/ST, large T antigen, small T antigen. **c**, Expression of REV-ERBs in BJ and BJ-ELR cells; quantitative PCR with reverse transcription (qRT–PCR), *n* = 3 biologically independent samples, two-tailed Mann–Whitney test. Expression of mRNA shown relative to housekeeper *RPLP0*. **d**, Jurkat cell viability is reduced by SR9009; *n* = 12 biological replicates, 72 h 20 µM, one-tailed Mann–Whitney test, *****P* < 0.0001. **e**, **f**, Immunostaining (**e**) and quantification (**f**) of cleaved caspase 3 (Cl. Casp. 3) and TUNEL assays (72 h, 20 µM); in **f**, *n* = 5 (mock) and 6 (SR9009) biologically independent samples, one-tailed Mann–Whitney test, cleaved caspase 3 assay, ***P* = 0.0022; TUNEL assay, ***P* = 0.0022. **g**, Breast cancer cell line MCF-7 viability is reduced by SR9009; *n* = 12 (mock) or 8 (SR9009) biological replicates, 72 h 20 µM, one-tailed Mann–Whitney test, *****P* < 0.0001. **h**, **i**, Immunostaining (**h**) and quantification (**i**) of cleaved caspase 3 and TUNEL assays (72 h, 20 µM). In **i**, *n* = 5 biologically independent samples, one-tailed Mann–Whitney test; cleaved caspase 3 assay, ***P* = 0.004; TUNEL assay, ***P* = 0.004. **j**, Colon cancer cell line HCT116 viability is reduced by SR9009; *n* = 8 biological replicates, water-soluble tetrazolium salt (WST-1) assay, 72 h, one-tailed Mann–Whitney test, *****P* < 0.0001. **k**, **l**, Induction of apoptosis is shown by cleaved caspase 3 and TUNEL staining (**k**, 72 h, 20 µM); for quantification in **l**, *n* = 8 (mock) or 5 (SR9009) biologically independent samples, one-tailed Mann–Whitney test; cleaved caspase 3 assay, ****P* = 0.0008; TUNEL assay ***P* = 0.0021. **m**–**o**, Prolonged SR9009 treatment eradicates cancer cells (7 days, 20 µM), but does not affect cells that express *NR1D1* and *NR1D2* shRNA. shNS, non-silencing shRNA. **p**, *NR1D1* and *NR1D2* qRT–PCR; *n* = 4 biologically independent samples; one-tailed Mann–Whitney test, **P* = 0.0286. NS, not significant. AU, arbitrary unit, shREV-ERBs, *NR1D1* and *NR1D2* shRNA. Scale bars, 50 µm. All panels representative of three biologically independent experiments. Data are mean ± s.e.m., except **c**, mean ± s.d.

desaturase 1 (SCD1)[14]. Unlike normal cells, cancer cells are highly dependent on *de novo* lipogenesis; major efforts are underway to develop cancer therapeutics on the basis of specific inhibitors of FAS and SCD1[19]. Agonists of REV-ERBs strongly reduced both mRNA and protein expression of these two key rate-limiting enzymes, which are involved in *de novo* lipogenesis (Extended Data Fig. 4a, b). This reduction led to the perturbation of several fatty acids and phospholipids (Extended Data Fig. 4c–i). Because oleic acid is the final product of SCD1 (Extended Data Fig. 4j), we investigated whether supplementing culture medium with oleic acid could attenuate the anticancer activity of agonists of REV-ERBs. Oleic acid impaired the anticancer activity of REV-ERB agonists (Extended Data Fig. 4k) but did not completely abrogate cytotoxicity, which suggests the involvement of additional mechanisms. By contrast, palmitic acid supplementation did not confer any protection (Extended Data Fig. 4l).

Cancer cells deal with their high metabolic demands through complex metabolic rewiring that involves the hyperactivation of autophagy[20]. Autophagy is essential for cancer cell survival, whereas normal cells depend on this catabolic cellular process only in starvation conditions[20]. Accordingly, inhibition of autophagy is a promising therapeutic strategy. However, chloroquine and its derivatives, which are the most common autophagy inhibitors, lack specificity and are toxic at high doses, potentially limiting their utility in clinical settings[21].

Autophagy is modulated in a circadian fashion and is controlled by NR1D1[15,22]. These observations prompted us to investigate whether the inhibition of autophagy is involved in the anticancer activity of agonists of REV-ERBs. We initially analysed the autophagosome marker LC3B to investigate whether agonists of REV-ERBs affect the numbers of autophagosomes; both SR9009 and SR9011 reduced the number of autophagosomes (Fig. 2a, b, Extended Data Fig. 5a, b). This decrease in autophagosomes suggests that the administration of agonists of REV-ERBs inhibited autophagy. To expand upon this observation, we tested whether p62 (also known as sequestosome-1), which is a protein that is specifically degraded by autophagy, accumulates following treatment with agonists of REV-ERBs. These agonists induced a marked accumulation of p62 in a range of cancer cell lines. (Fig. 2c–e, Extended Data Fig. 5c–e). If autophagy has a dominant role in the induction of apoptosis triggered by agonists of REV-ERBs, autophagy inhibition should precede apoptosis induction: indeed, the blockage of autophagy shown by p62 accumulation occurred before the induction of apoptosis (Fig. 2f, g, Extended Data Fig. 5f, g). The inhibition of autophagy was further confirmed by analysis of the autophagic flux and by electron microscopy, with the latter showing that autophagosome formation was also impaired on starvation (Extended Data Fig. 6a–c). In addition, treatment with agonists of REV-ERBs prevented lysosome turnover, as shown by an increase in the lysosomal protein LAMP1 and by the enhanced activity of LysoTracker Red, which stains acidic vesicles (Extended Data Fig. 6d, e). The accumulation of lysosomes was also observed by transmission electron microscopy (Extended Data Fig. 6f). Together, these results indicate that agonists of REV-ERBs potently inhibit autophagy.

When challenged with starvation, cancer cells are extremely sensitive to the inhibition of autophagy. The cytotoxicity of SR9009 and SR9011 was enhanced by starvation in a range of cancer cell lines (Fig. 2h, Extended Data Fig. 6g, h), which indicates the involvement of autophagy inhibition. Starvation did not induce the expression of REV-ERBs (Extended Data Fig. 6i, j), showing that autophagy inhibition is responsible for the increased sensitivity to agonists of REV-ERBs in starved cancer cells. Finally, the overexpression of core autophagy genes (*ULK2*, *ULK3* and *LKB1* (also known as *STK11*) abrogated

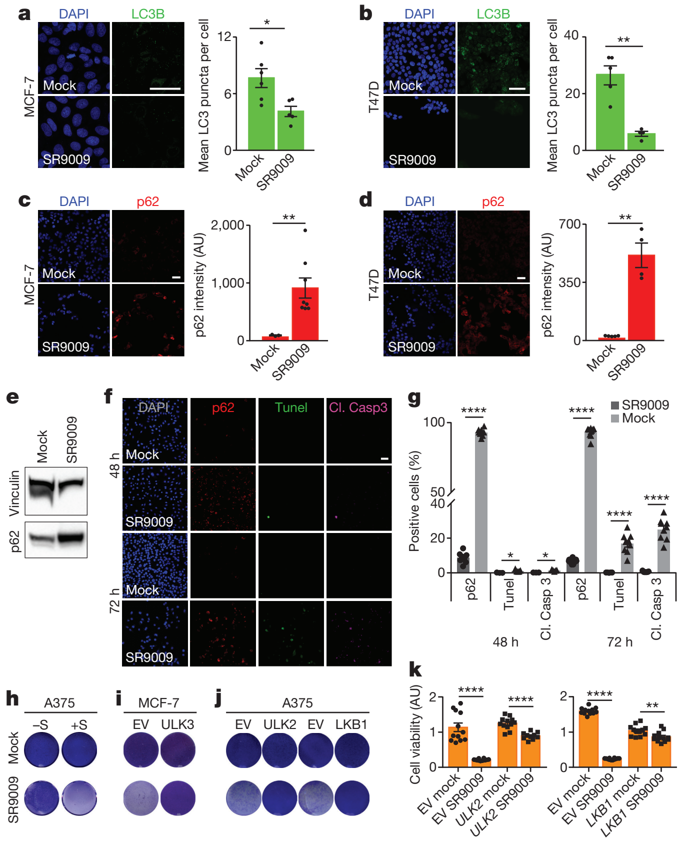


Figure 2 | SR9009 agonist of REV-ERBs inhibits autophagy. **a**, **b**, SR9009 treatment reduces the number of autophagosomes, as shown by immunofluorescence of LC3B. *n* indicates biologically independent samples. MCF-7, $n = 6$ (mock) or 5 (SR9009); breast cancer cell line T47D, $n = 5$ (mock) or 4 (SR9009). One-tailed Mann–Whitney test; MCF-7 20 μM 24 h, $*P = 0.0152$, T47D 20 μM 48 h, $**P = 0.0079$. **c**, **d**, SR9009 induces accumulation of p62 as shown by immunofluorescence. *n* indicates biologically independent samples; MCF-7, $n = 3$ (mock) or 8 (SR9009); T47D, $n = 5$ (mock) or 4 (SR9009). One-tailed Mann–Whitney test; MCF-7 p62 48 h, $**P = 0.0061$; T47D 48 h, $**P = 0.0079$. **e**, Inhibition of autophagy is confirmed by the immunoblot for p62 (20 μM, 48 h, melanoma cell line A375). **f**, **g**, Inhibition of autophagy precedes apoptosis induction as shown by immunofluorescence of p62, cleaved caspase 3 and TUNEL assays. *n* indicates biologically independent samples. One-tailed Mann–Whitney test; A375 20 μM, cleaved caspase 3 assay 48 h, $n = 3$, $*P = 0.0179$; cleaved caspase 3 assay 72 h, $n = 7$, $****P < 0.0001$; TUNEL assay 48 h, $n = 3$, $*P = 0.0179$; TUNEL assay 72 h, $n = 7$, $****P < 0.0001$; p62 48 h, $n = 8$, $****P < 0.0001$; p62 72 h, $n = 9$, $****P < 0.0001$. **h**, Starvation markedly accelerates the cytotoxic effect of the REV-ERB agonist SR9009 (A375, 3 days, 20 μM, starvation time 24 h). **i**, Overexpression of *ULK3* impairs SR9009 induction of apoptosis (MCF-7, 6 days, 20 μM). **j**, Overexpression of *ULK2* and *LKB1* impairs SR9009 induction of apoptosis (A375, 6 days, 20 μM). **k**, WST-1 viability assay shows abrogation of apoptosis in *ULK2* (left panel) and *LKB1* (right panel) overexpressing cells (6 days). *n* indicates biological replicates. One-tailed Mann–Whitney test; left panel A375, 20 μM, $n = 12$, empty vector (EV) mock versus EV SR9009, $****P < 0.0001$; *ULK2* mock ($n = 12$) versus *ULK2* SR9009 ($n = 10$), $****P < 0.0001$; right panel A375, $n = 12$, EV mock versus EV SR9009, $****P < 0.0001$; *LKB1* mock versus *LKB1* SR9009, $**P = 0.0028$. Scale bars, 50 μm. All panels representative of three biologically independent experiments with similar results. All data are mean ± s.e.m. For gel source data, see Supplementary Fig. 1.

induction of apoptosis in various tumour cell lines (Fig. 2i–k, Extended Data Fig. 6k–p).

Next, we sought to determine how agonists of REV-ERBs block autophagy. We initially compared differential autophagy outcomes observed between chloroquine and agonists of REV-ERBs. Chloroquine inhibits autophagy at a late stage by blocking the fusion of autophagosomes and lysosomes, and thereby leads to the accumulation of autophagosomes (Extended Data Fig. 6a–c). By contrast, agonists of REV-ERBs decreased the number of autophagosomes, which suggests that they block autophagy at an early stage.

To gain additional mechanistic insights, we investigated whether agonists of REV-ERBs can regulate the expression of core autophagy genes. Analysis of a published report[5] on chromatin occupancy of REV-ERBs revealed the presence of peaks in *Ulk3*, *Ulk1*, *Becn1* and *Atg7* (Extended Data Fig. 7a). Using HOMER software (http://homer.ucsd.edu/homer/), we found that NR1D1 and NR1D2 consensus binding sites were also present within these genetic loci (Extended Data Fig. 7b–e). Accordingly, *ULK3*, *ULK1*, *BECN1* and *ATG7* mRNAs and protein levels of ULK3, ULK1, BECN1 and ATG7 were downregulated on treatment with agonists of REV-ERBs, whereas the expression of these genes was induced following depletion of REV-ERBs by shRNA (Extended Data Figs 7f–j, 8a–e). Furthermore, in REV-ERB-depleted cells, agonists of REV-ERBs did not repress autophagy genes (Extended Data Figs 7k, 8f).

The activation of aberrant oncogenic stimuli is an early step in tumorigenesis. Oncogene-induced senescence (OIS) arises in normal cells to limit the expansion of cells affected by oncogenic stress[23,24]. Although this provides an immediate benefit by arresting potentially dangerous cells, the accumulation of senescent cells over long periods of time contributes to tumour formation, tumour progression and age-related diseases[25]. Furthermore, the induction of cellular senescence upon anticancer chemotherapy treatment promotes chemotherapy resistance and generates an environment that may support uncontrolled growth of neighbouring cells and fuel relapse[25]; this highlights the need for senolytic agents. Although *de novo* lipogenesis is upregulated in cancer cells but not in OIS cells[26], an elevated level of autophagy is a known characteristic of OIS cells[27]. Agonists of REV-ERBs are lethal when administered to cells characterized by oncogenic RAS signalling (Fig. 1, Extended Data Fig. 1), affect slowly proliferating cancer stem cells and potently inhibit autophagy (Fig. 2, Extended Data Figs 5, 6, 8g–l).

We investigated whether treatment of OIS cells with agonists of REV-ERBs would block autophagy and lead to apoptosis. The overexpression of the HRAS proto-oncogene GTPase with the oncogenic mutation G12V (HRAS$^{\mathrm{G12V}}$) (Extended Data Fig. 9a) established OIS, as shown by an increase in senescence-associated β-galactosidase activity and by upregulation of cell-cycle inhibitors (Extended Data Fig. 9b, c). Notably, agonists of REV-ERBs triggered the induction of apoptosis in OIS cells without affecting normal proliferating or quiescent cells (Fig. 3a–c, Extended Data Fig. 9d–g). In agreement with previous results (Fig. 2c–e, Extended Data Fig. 5c–e), treatment with agonists of REV-ERBs led to the accumulation of p62 and lysosomes and a reduction in autophagosomes (Fig. 3d, e, Extended Data Fig. 9h, i). Therefore, agonists of REV-ERBs inhibit autophagy in OIS cells. Finally, on overexpression of ULK3, the pro-apoptotic ability of the agonists of REV-ERBs was impaired (Fig. 3f). These results show that through their ability to block autophagy, these agonists can target a premalignant non-proliferating cellular population. Therefore, agonists of REV-ERBs display senolytic activity in addition to their oncolytic effects.

To understand whether agonists of REV-ERBs represent an effective therapeutic strategy, we investigated whether agonists of REV-ERBs affect OIS and tumour viability *in vivo*. Naevi are benign lesions consisting of cutaneous melanocytes that have undergone OIS upon aberrant activation of RAS signalling[28]. Consistent with the *in vitro* observations discussed earlier, SR9009 treatment led to an increase in apoptosis in NRAS-induced naevi in mice, and the repression of autophagy genes (Fig. 4a–c, Extended Data Fig. 10a). This indicates the potential for non-proliferating premalignant cells to be selectively targeted by agonists of REV-ERBs *in vivo*. However, although naevi have been used as a model for studying OIS *in vivo*[28], they do not

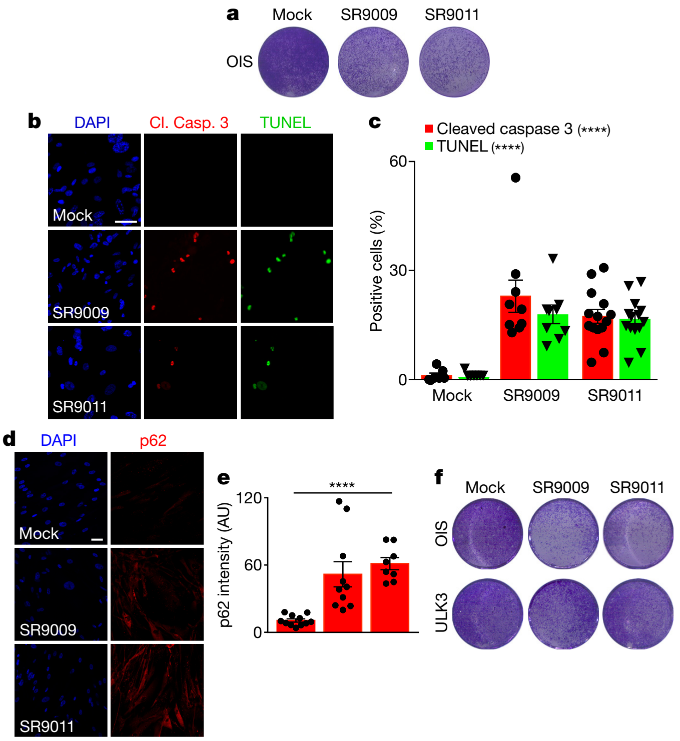


Figure 3 | SR9009 and SR9011 treatment evokes an apoptotic response and induces inhibition of autophagy in OIS cells. a, Proliferation assay shows that agonists of REV-ERBs impair viability of OIS cells (6 days, 20 μM). **b**, **c**, Immunofluorescence assay for cleaved caspase 3 and TUNEL assay show apoptosis specifically in OIS cells. *n* indicates biologically independent samples; $n = 7$ (mock), $n = 9$ (SR9009) and $n = 14$ (SR9011). One-way ANOVA, 72 h, 20 μM; cleaved caspase 3 assay, $^{****}P < 0.0001$; TUNEL assay, $^{****}P < 0.0001$; mean ± s.e.m. **d**, **e**, p62 accumulates on treatment with agonists of REV-ERBs, as assayed by immunofluorescence for p62. *n* indicates biologically independent samples, $n = 11$ (mock), $n = 10$ (SR9009) and $n = 8$ (SR9011). One-way ANOVA, 72 h, 20 μM, $^{****}P < 0.0001$; mean ± s.e.m. **f**, ULK3 overexpression protects OIS cells from cytotoxicity induced by agonists of REV-ERBs; 6 days, 20 μM. Scale bars, 50 μm. All panels representative of three biologically independent experiments with similar results.

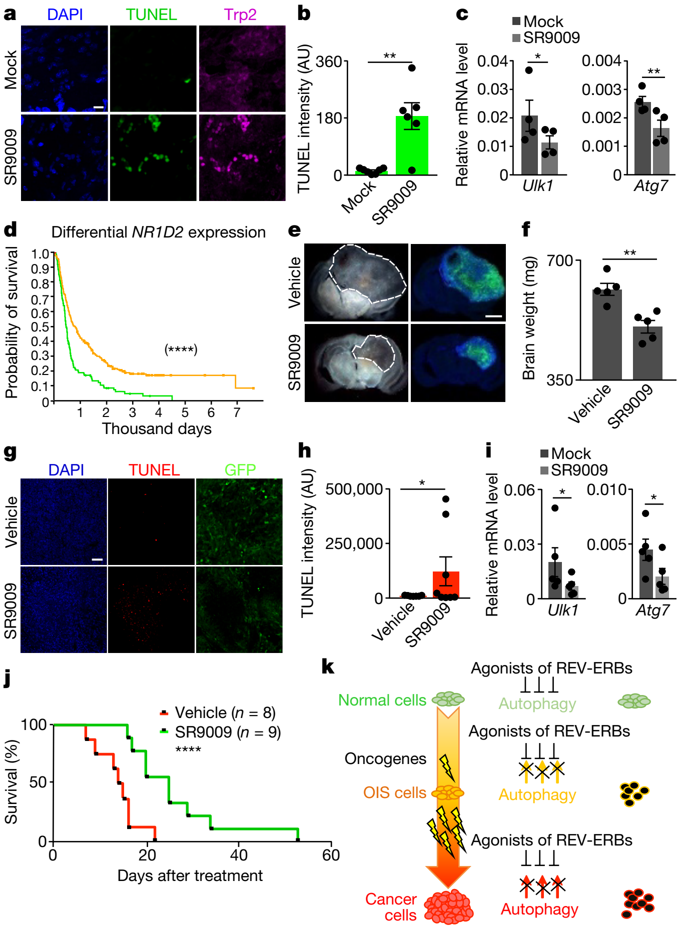


Figure 4 | SR9009 impairs viability of NRAS-driven naevi and glioblastoma growth and extends survival. a, **b**, SR9009 treatment induces apoptosis *in vivo* in NRAS-driven naevi, as assayed by immunofluorescence analysis (representative images of two independent experiments with similar results, TRP2 melanocytic marker and TUNEL assay; one-tailed Mann–Whitney test, $^{**}P = 0.0058$. *n* indicates biologically independent samples, $n = 7$ (mock) and 6 (SR9009), 12 days of SR9009, 20 μM, four mice. Scale bar, 10 μm. **c**, Autophagy genes are downregulated after treatment of NRAS-driven naevi, $n = 4$ mice. One-tailed Mann–Whitney; *Ulk1*, $^{*}P = 0.0249$ and *Atg7*, $^{**}P = 0.007$. **d**, *NR1D2* expression correlates with survival in patients with brain cancer. *n* indicates biologically independent samples. Yellow line, intermediate expression ($n = 224$); green line, downregulated ($n = 119$). NIH Rembrandt database (https://wiki.nci.nih.gov/display/ICR/Rembrandt+Data+Portal); two-sided log-rank, $^{****}P < 0.0001$. **e**, **f**, SR9009 treatment impairs *in vivo* growth of glioblastoma. Representative images of one experiment, $n = 5$ mice, 6 days, 200 mg kg^{-1} SR9009 twice a day. One-tailed Mann–Whitney test, $^{**}P = 0.004$. **g**, **h**, SR9009 induces apoptosis in glioblastoma, as shown by TUNEL assay; tumour cells are GFP-positive. Representative images of one independent experiment, 6 days 200 mg kg^{-1} twice a day. One-tailed Mann–Whitney test, $^{*}P = 0.02$. *n* indicates biologically independent samples, $n = 7$ (mock) or 8 (SR9009), five mice. **i**, *In vivo* treatment with SR9009 results in downregulation of main autophagy genes. Six days, 200 mg kg^{-1} twice a day, $n = 5$ mice. One-tailed Mann–Whitney test, $^{*}P = 0.0476$. **j**, SR9009 improves survival of mice affected by glioblastoma. SR9009, 100 mg kg^{-1}. Vehicle, $n = 8$; SR9009, $n = 9$ mice; two-tailed log-rank, $^{***}P = 0.0009$. **k**, Scheme illustrating how agonists of REV-ERBs selectively affect OIS and cancer cells. All bar charts mean ± s.e.m.

affect neighbouring tissues and do not develop into melanomas: further studies are necessary to assess the therapeutic relevance of agonists of REV-ERBs as senolytic tools[25,29]. As previously reported[14,15], agonists of REV-ERBs do not show overt toxicity; our TUNEL analyses—performed in normal skin and brain tissues—and our body weight analyses confirm this finding (Extended Data Fig. 10b–d). By contrast, established anticancer agents such as temozolomide are characterized by several side effects (Extended Data Fig. 10d).

SR9009 is known to cross the blood–brain barrier[14], and several glioblastoma cell lines—including brain-tumour initiating cells (005 and RIGH), A172 and glioblastoma stem cells derived from patients—are sensitive to treatment with agonists of REV-ERBs *in vitro* (Fig. 1a, Extended Data Figs 1a, 2f, g, 8g and 10e). On analysis of REV-ERB status in glioblastoma data from The Cancer Genome Atlas (http://www.cbioportal.org/)[30,31], we observed that NR1D1 and NR1D2 status was unaltered in nearly all patients with glioblastoma (Extended Data Fig. 10f, and data not shown), which suggests a possible therapeutic use for agonists of REV-ERBs in clinical settings.

Finally, NR1D2 expression was positively correlated with survival in brain cancer patients (Fig. 4d, NR1D1 data are not available). For the above reasons, and because of the low toxicity of agonists of REV-ERBs, we tested SR9009 for treating brain tumours. Brain-tumour-initiating cells were transplanted into mouse brains by stereotaxic injection, and on tumour establishment, SR9009 treatment was initiated. SR9009 reduced glioblastoma growth, triggered apoptosis and downregulated the expression of autophagy genes (Fig. 4e–i, Extended Data Fig. 10g). Additionally, SR9009 reduced tumour growth in a xenograft model of a patient-derived glioblastoma (Extended Data Fig. 10h, i). Most notably, SR9009 effectively and significantly improved survival in two glioblastoma models, including a xenograft derived from a patient (Fig. 4j, Extended Data Fig. 10j). The anticancer activity of SR9009 was similar to that of temozolomide, the current therapeutic standard for glioblastoma (Extended Data Fig. 10j). Unlike temozolomide, however, SR9009 lacked toxicity.

Together, these results indicate that agonists of REV-ERBs are pharmacological tools that target a potentially wide spectrum of tumours and therapeutic windows, are highly selective and exhibit low toxicity (Fig. 4k). Importantly, REV-ERB agonists selectively target slowly proliferating tumorigenic and premalignant populations, such as naevi (Fig. 4k). We propose that the anticancer activity of agonists of REV-ERBs involves the inactivation of at least two cancer hallmarks[3]: *de novo* lipogenesis and autophagy, with autophagy—given its importance in meeting the metabolic demands of cancer cells—possibly having a major role. By simultaneously targeting two essential cancer hallmarks, agonists of REV-ERBs may represent an improved therapeutic strategy for treating cancer. These results strongly indicate that pharmacological modulation of circadian machinery is an innovative and selective strategy for cancer treatment (Fig. 4k).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 September 2016; accepted 24 November 2017.
Published online 10 January 2018.

1. Fu, L. & Lee, C. C. The circadian clock: pacemaker and tumour suppressor. *Nat. Rev. Cancer* **3,** 350–361 (2003).
2. Scheiermann, C., Kunisaki, Y. & Frenette, P. S. Circadian control of the immune system. *Nat. Rev. Immunol.* **13,** 190–198 (2013).
3. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
4. Straif, K. *et al.* Carcinogenicity of shift-work, painting, and fire-fighting. *Lancet Oncol.* **8,** 1065–1066 (2007).
5. Cho, H. *et al.* Regulation of circadian behaviour and metabolism by REV-ERB-α and REV-ERB-β. *Nature* **485,** 123–127 (2012).
6. Bugge, A. *et al.* REV-ERBα and REV-ERBβ coordinately protect the circadian clock and normal metabolic function. *Genes Dev.* **26,** 657–667 (2012).
7. Yu, E. A. & Weaver, D. R. Disrupting the circadian clock: gene-specific effects on aging, cancer, and other phenotypes. *Aging* **3,** 479–493 (2011).
8. Plikus, M. V. *et al.* Local circadian clock gates cell cycle progression of transient amplifying cells during regenerative hair cycling. *Proc. Natl Acad. Sci. USA* **110,** E2106–E2115 (2013).
9. Fu, L., Pelicano, H., Liu, J., Huang, P. & Lee, C. The circadian gene *Period2* plays an important role in tumor suppression and DNA damage response *in vivo*. *Cell* **111,** 41–50 (2002).
10. Sancar, A. *et al.* Circadian clock control of the cellular response to DNA damage. *FEBS Lett.* **584,** 2618–2625 (2010).
11. Bass, J. Circadian topology of metabolism. *Nature* **491,** 348–356 (2012).
12. Preitner, N. *et al.* The orphan nuclear receptor REV-ERBα controls circadian transcription within the positive limb of the mammalian circadian oscillator. *Cell* **110,** 251–260 (2002).
13. Yin, L. *et al.* REV-ERBα, a heme sensor that coordinates metabolic and circadian pathways. *Science* **318,** 1786–1789 (2007).
14. Solt, L. A. *et al.* Regulation of circadian behaviour and metabolism by synthetic REV-ERB agonists. *Nature* **485,** 62–68 (2012).
15. Woldt, E. *et al.* REV-ERB-α modulates skeletal muscle oxidative capacity by regulating mitochondrial biogenesis and autophagy. *Nat. Med.* **19,** 1039–1046 (2013).
16. Vieira, E. *et al.* The clock gene *Rev-erbα* regulates pancreatic β-cell function: modulation by leptin and high-fat diet. *Endocrinology* **153,** 592–601 (2012).
17. Gorrini, C., Harris, I. S. & Mak, T. W. Modulation of oxidative stress as an anticancer strategy. *Nat. Rev. Drug Discov.* **12,** 931–947 (2013).
18. Peek, C. B. *et al.* Circadian clock NAD$^+$ cycle drives mitochondrial oxidative metabolism in mice. *Science* **342,** 1243417 (2013).
19. Currie, E., Schulze, A., Zechner, R., Walther, T. C. & Farese, R. V. Jr. Cellular fatty acid metabolism and cancer. *Cell Metab.* **18,** 153–161 (2013).
20. White, E. Deconvoluting the context-dependent role for autophagy in cancer. *Nat. Rev. Cancer* **12,** 401–410 (2012).
21. Rubinsztein, D. C., Codogno, P. & Levine, B. Autophagy modulation as a potential therapeutic target for diverse diseases. *Nat. Rev. Drug Discov.* **11,** 709–730 (2012).
22. Ma, D., Panda, S. & Lin, J. D. Temporal orchestration of circadian autophagy rhythm by C/EBPβ. *EMBO J.* **30,** 4642–4651 (2011).
23. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic *ras* provokes premature cell senescence associated with accumulation of p53 and p16^{INK4a}. *Cell* **88,** 593–602 (1997).
24. Campisi, J. & d'Adda di Fagagna, F. Cellular senescence: when bad things happen to good cells. *Nat. Rev. Mol. Cell Biol.* **8,** 729–740 (2007).
25. Childs, B. G. *et al.* Senescent cells: an emerging target for diseases of ageing. *Nat. Rev. Drug Discov.* **16,** 718–735 (2017).
26. Quijano, C. *et al.* Oncogene-induced senescence results in marked metabolic and bioenergetic alterations. *Cell Cycle* **11,** 1383–1392 (2012).
27. Young, A. R. *et al.* Autophagy mediates the mitotic senescence transition. *Genes Dev.* **23,** 798–803 (2009).
28. Michaloglou, C. *et al.* BRAF^{E600}-associated senescence-like cell cycle arrest of human naevi. *Nature* **436,** 720–724 (2005).
29. Chang, J. *et al.* Clearance of senescent cells by ABT263 rejuvenates aged hematopoietic stem cells in mice. *Nat. Med.* **22,** 78–83 (2016).
30. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155,** 462–477 (2013).
31. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank K. V. Ly, L. Fijany, Y. Soda, M. Soda and M. Schmitt for technical assistance; F. d'Adda di Fagagna, S. Minucci, A. Viale, G. Gargiulo and J. Karlseder for discussion and feedback; the Narita, Gage, Burris, Amati and Shaw laboratories, and F. F. Lang for reagents; and the Salk Institute's Waitt Advanced Biophotonics Center and Gene Targeting and Transfer, and M. Shokhirev and the Razavi Newman Integrative Genomics and Bioinformatics Core. G.S. is supported by the AIRC/Marie Curie International Fellowships in Cancer Research (12298), Istituto Superiore di Sanità, TRAIN 'Training through Research Application Italian iNitiative'. M.V.P. is supported by the NIH (NIAMS grants R01-AR067273 and R01-AR069653) and a Pew Charitable Trust grant. X.W. is supported by a CIHR postdoctoral fellowship (MFE-123724). I.M.V is an American Cancer Society Professor of Molecular Biology. A.R. is supported by the NCI T32 grant, Salk Women in Science, Salk Excellerators Award and the Stavros Niarchos Foundation New Frontiers Salk Research Specialist Award. M.J.K. is supported by F30 DK112604. A.S. is supported by the NCI Cancer Center Support Grant P30 (CA014195 MASS core) and Dr. Frederick Paulsen Chair/Ferring Pharmaceuticals. This work was supported in part by a Worldwide Cancer Research grant and an American Federation of Aging Research mid-career grant M14322 to S.P. Additional support came from a Cancer Center Core Grant (P30 CA014195-38), the H. N. and Frances C. Berger Foundation, the Glenn Center for Aging Research and the Leona M. and Harry B. Helmsley Charitable Trust (grant #2012-PG-MED002).

Author Contributions X.W. and M.V.P. performed experiments related to naevi. A.R. participated in xenograft experiments. F.P. performed experiments in Extended Data Figs 8g–l, 10h–j. M.J.K. performed lipidomic assays. G.S. designed the study, performed experiments, analysed data and wrote the manuscript. A.S, I.M.V. and M.V.P. supervised experiments and edited the manuscript. S.P. supervised, designed the study and reviewed the data and manuscript. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to G.S. (gsulli@salk.edu) and S.P. (satchin@salk.edu).

Reviewer Information *Nature* thanks S. Kay, L. Zender and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Cell culture and treatments. BJ, WI38, BJ-ELR, A375, Jurkat, MCF7, T47D, HCT116, Becker (astrocytoma line, JCRB Cell Bank), PANC-1 and SK-MEL28 cells were grown under standard tissue-culture conditions and obtained through ATCC. No further authentication was performed. HCT116 $p53^{-/-}$ cells were a gift from B. Amati. BTICs (005, RIGH) were cultured as previously described[32]. OIS cells were generated as previously described[33]. Quiescent cells were obtained by contact inhibition. Cell lines were tested for mycoplasma contamination. Senescence-associated (SA)-β-galactosidase assay (Cell Biolabs) was performed as previously described[33]. SR9009 (Xcessbio, Millipore) and SR9011 (Xcessbio) were dissolved in DMSO for *in vitro* studies and for ear topical administration; SR9009 and temozolomide (Cayman Chemicals) were dissolved in 15% Cremophor (Sigma-Aldrich) in sterile water for *in vivo* studies. Hypoxia was induced by lowering incubator oxygen percentage to 1–2%, or with NAC supplementation in the medium (10 mM; Sigma-Aldrich); EBSS (Life Technologies) was used to induce starvation. Proliferation assays were performed to assess the cytotoxicity of SR9009 and SR9011 in normal and cancer cells by using crystal violet and cell proliferation reagent WST-1 (Roche); all treatment started when cells were 80% confluent (except for the BTIC experiments). MTS (3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium) assays were performed according to the manufacturer's instructions (CellTiter 96 Aqueous One, Promega).

Human samples. Glioblastoma stem cells (GSCs) were isolated from specimens from patients with glioblastoma, who had undergone surgery at the University of Texas MD Anderson Cancer Center (UTMDACC)[34]. The diagnosis of glioblastoma was established by histological examination, according to the WHO (World Health Organization) classification. Samples derived from patients were obtained with the consent of patients, under an ethically approved Institutional Review Board protocol LAB04-0001 chaired by F. F. Lang (UTMDACC). Tumour specimens were dissociated and resuspended in Dulbecco's modified Eagle's medium/F12 (Gibco) supplemented with B27 (×1, Invitrogen), bFGF (20 ng ml^{-1} Peprotech), and EGF (20 ng ml^{-1}, Peprotech). Cells were cultured as neurospheres and passaged every 5–7 days, on the basis of sphere size.

Plasmids. pBABE-Puro and pBABE-Puro H-RasV12 were used as previously described[33]. pLKO.1 *NR1D1* shRNA (shNR1D1), pLKO.1 *NR1D2* shRNA (shNR1D2) (Sigma-Aldrich), pLPCULK3, pLPCLC3B (gift from the Narita laboratory) pLenti-ULK2 (ABM), pBABE-LKB1 (gift from the Shaw laboratory) were obtained as indicated. shNR1D1 no. 1: CCGGGCGCTTTGCTTCGTTGTTCAACTCGAGTTGAACAACGAAGCAAAGCGCTTTTT; shNR1D1 no. 2: CCGGCCAGCCCTGAATCCCTCTATACTCGAGTATAGAGGGATTCAGGGCTGGTTTTT; shNR1D2 no. 1: CCGGGCCCTCCAACTTAGTGATGAACTCGAGTTCATCACTAAGTTGGAGGGCTTTTT; shNR1D2 no. 2: CCGGCCAGTACAAGAAGTGCCTGAACTCGAGTTCAGGCACTTCTTGTACTGGTTTTT.

Immunofluorescence microscopy. For brain-tissue immunofluorescence, all mice were perfused with 0.9% NaCl followed by 4% paraformaldehyde in PBS. The brains were collected, fixed overnight and transferred to 30% sucrose in PBS. For fluorescent staining, 40-μm coronal sections on a sliding microtome were prepared and imaged with the Zeiss LSM 780 Side Port Laser Scanning Confocal microscope. Mouse ears were fixed in 4% paraformaldehyde and subjected to histology. Paraffin-embedded sections were stained with anti-TRP2 at 2 μg ml^{-1} (Santa Cruz), TUNEL *In Situ* Cell Death Detection Kit, Fluorescein (Roche). Cells were fixed and probed as previously described[33]. Images and confocal sections were acquired using the Zeiss LSM 780 Side Port Laser Scanning Confocal microscope. Comparative immunofluorescence microscopy analyses were performed in parallel, with identical acquisition and analysis parameters. ImageJ software (v1.49g) was used to perform quantitative analyses and to assay intensity differences. To count LC3B puncta, after selecting a threshold to minimize any effect of background signal analyses were performed on projected stack using the ImageJ function 'analyse particles'. 3D Objects Counter was used to analyse intensity. Apoptosis was evaluated by the immunostaining of cleaved caspase 3 (Cell Signaling No. 9664 1:200), and by TUNEL assay using *In Situ* Cell Death Detection Kit, Fluorescein or TMR red (Roche). Antibodies used were LC3B (Cell Signaling No. 3868 1:200), Lamp1 (Cell Signaling #9091 1:200), Sqstm1/p62 (Abcam ab56416 1:100), Sqstm1/p62 (MBL PM045) and Ras (BD #610002, 1:200). LysoTracker Red DND-99 (Lifetech) was used to visualize lysosomes.

Electron microscopy. Cells grown on ACLAR coverslips were fixed in 2.5% glutaraldehyde with 2% paraformaldehyde in 0.15 M cacodylate buffer containing 2 mM calcium chloride, pH 7.4, at 37 °C for five minutes, followed by an hour at 4 °C. The coverslips were then washed in 0.15 M cacodylate buffer containing 2 mM calcium chloride, and secondarily fixed in 1% osmium tetroxide and 0.3% potassium ferrocyanide in the same buffer. Subsequently, the coverslips were washed in water and stained en bloc with 2% uranyl acetate, followed by a graded dehydration in ethanol (35%, 50%, 70%, 90%, 100% and 100%). Samples were then rapidly infiltrated in EPON resin using a Pelco BioWave microwave processing unit (Ted Pella), flat embedded and cured at 60 °C overnight. Regions of interest were excised and remounted on blank resin stubs. Ultrathin (70 nm) sections were then cut on a UC7 ultramicrotome (Leica) and cells were imaged on a transmission electron microscope at 120 kV (Zeiss Libra 120 PLUS).

Immunoblotting. Cells were lysed in sample buffer and 20–50 μg of whole cell lysate was resolved by gel electrophoresis (Bolt 4–12% Bis-Tris Plus Gels, Life Technologies), transferred to nitrocellulose (iBlot Transfer Stack, nitrocellulose, Life Technologies) and probed with the following antibodies: anti-cleaved caspase 3 (1:250 Cell Signaling No. 9664); anti-vinculin clone hVIN-1 (SIGMA; 1:10,000), anti Sqstm1/p62 (Abcam ab56416; 1:500), BECN1 (Santacruz H-300, sc-11427; 1:250), ATG7 (Sigma-Aldrich, A2856; 1:1,000), ULK1 (Sigma-Aldrich, A7481; 1:250), ULK3 (Sigma-Aldrich, SAB4200132; 1:500), SCD1 (Abcam, ab19862; 1:1,000), FASN (Cell Signaling, 3180; 1:1,000) and Tubulin (Millipore, 05-829; 1:5,000).

qRT–PCR. Total RNA was extracted from cells and tissues using RNAeasy (Qiagen) according to the manufacturer's instructions, and treated with DNase before reverse transcription. cDNA was generated using qScript cDNA SuperMix (Quanta BioSciences). The cDNA was used as a template in real-time quantitative PCR reactions with specific primers on an ABI 7900HT Fast Real-Time PCR System. The reactions were prepared using SyBR Green reaction mix from Roche. The gene (*RPLP0*) encoding ribosomal protein P0 (RPP0) was used as a control for normalization. Human primer sequences for qRT–PCR: *RPLP0*-fw, 5′-TTCATTGTGGGAGCAGAC-3′; *RPLP0*-rev, 5′-CAGCAGTTTCTCCAGAGC-3′; *NR1D1*-fw, 5′-GCATGGAGAATTCCGCTTC-3′; *NR1D1*-rev, 5′-CGGTTCTTCAGCACCAGAG-3′; *NR1D2*-fw, 5′-CATTTCTATATTTGAAAGTAGCCCAAT-3′; *NR1D2*-rev, 5′-ACTCAATCAAAGAATGTGCTTGTAA-3′; *ULK2*-fw, 5′-TCAAGCATCTTCCAACCTGTT-3′; *ULK2*-rev, 5′-ATTCCCGTGGCTCATTCCCAT-3′; L*KB1*-fw, 5′-GAGCTGATGTCGGTGGGTATG-3′; *LKB1*-rev, 5′-CACCTTGCCGTAAGAGCCT-3′; *ULK1*-fw, 5′-AAGCACGATTTGGAGGTCGC-3′; *ULK1*-rev, 5′-TGATTTCCTTCCCCAGCAGC-3′; *BECN1*-fw, 5′-CCATGCAGGTGAGCTTCGT-3′; *BECN1*-rev, 5′-GAATCTGCGAGAGACACCATC-3′; *ULK3*-fw, 5′-TGAAGGAGCAGGTCAAGATGA-3′; *ULK3*-rev, 5′-GCTACGAACAGATTCCGACAG-3′; *CDKN2B*-fw, 5′-GCGGGGACTAGTGGAGAAG-3′; *CDKN2B*-rev, 5′-CTGCCCATCATCATGACCT-3′; *CDKN2A*-fw, 5′-CCCAACGCACCGAATAGTTAC-3′; *CDKN2A*-rev, ATTCCAATTCCCCTGCAAACT-3′; *SCD1*-fw, 5′-GACGATGAGCTCCTGCTGTT-3′; *SCD1*-rev, 5′-CTCTGCTACACTTGGGAGCC-3′; *FASN*-fw, 5′-CATCGGCTCCACCAAGTC-3′; *FASN*-rev, 5′-GCTATGGAAGTGCAGGTTGG-3′; Mouse primer sequences for qRT–PCR: *ulk1*-fw, 5′-GAGCCGAGAGTGGGGCTTTGC-3′; *ulk1*-rev, 5′-GCCCTGGCAGGATACCACGC-3′; *atg7*-fw, 5′-CCGGTGGCTTCCTACTGTTA-3′; *atg7*-rev, 5′-AAGGCAGCGTTGATGACC-3′.

Chromatin immunoprecipitation with sequencing data analysis. Peak calling was performed using model-based analysis of chromatin immunoprecipitation with sequencing (ChIP–seq) (MACS)[35] with Galaxy Tool Version 1.0.1[35]. *P*-value cutoff for peak detection was selected as $\leq 10^{-5}$, and the false discovery rate as ≤ 0.05.

Promoter motif analysis. Regions that are 2,000 bp upstream and 100 bp downstream from the transcription start sites of *Becn1*, *Atg7*, *Ulk1* and *Ulk3* were scanned using the mouse (mm10) genome annotation using the HOMER v4.9.1 findMotifs.pl script with start = −2,000 and end = 100 and a log odds-score threshold of 5, looking for the motif 'GTAGGTCACTGGGTCA' and trained on data from a previous study[36].

Lipid extraction. Lipid extraction was performed as previously described[37,38]. A172 cells were vortexed for 30 s in a mixture of 1:1:2 PBS:methanol:chloroform. A $^{13}C_{16}$-palmitic acid standard (200 pmol per sample) was added to chloroform before extraction. The resulting mixture was centrifuged at 2,200 *g*, 5 min, 4 °C to separate organic and aqueous phases. The organic phase (bottom layer) was collected and dried under a stream of nitrogen.

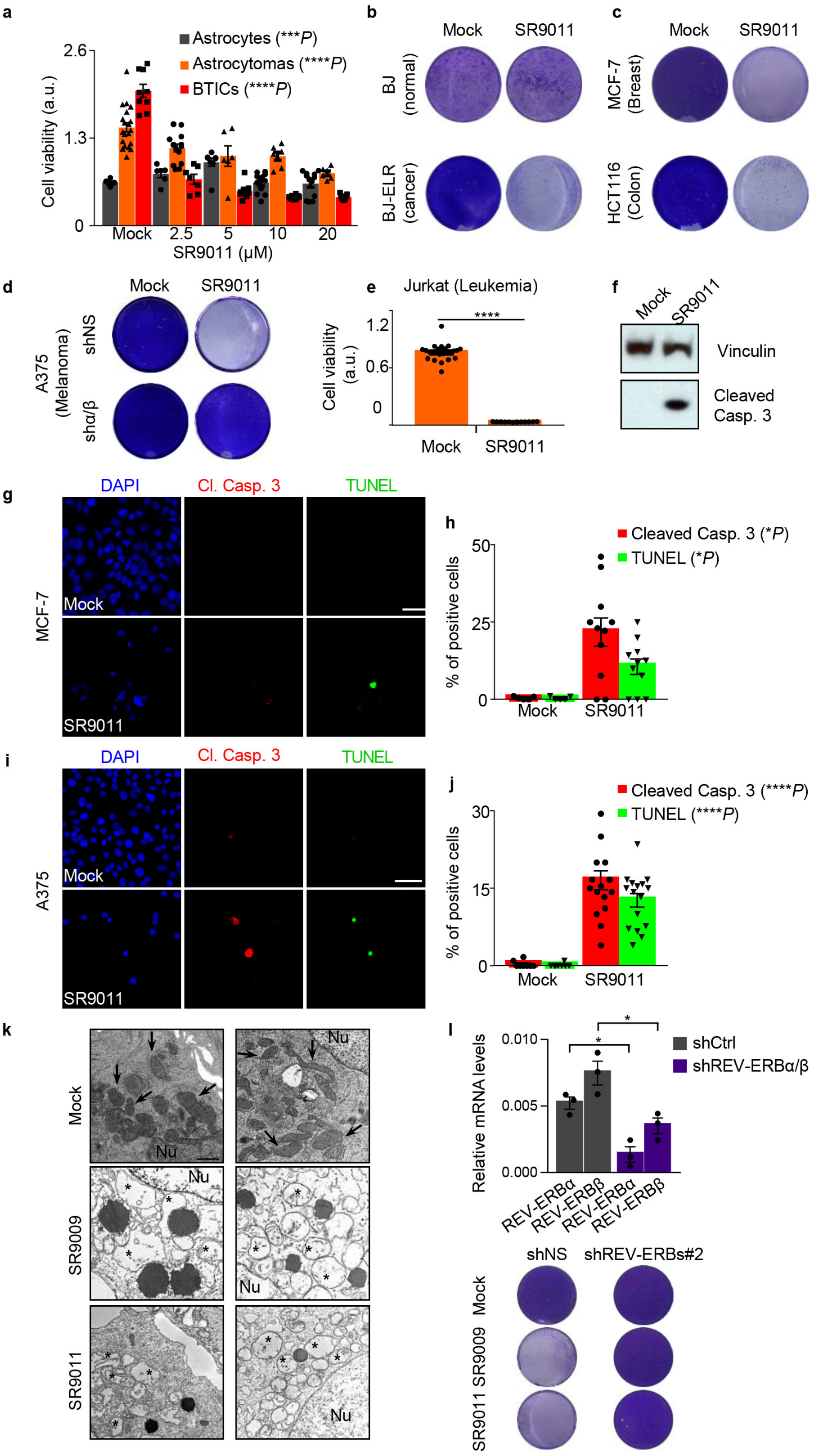
Lipidomic analysis. Lipidomic analysis was performed as previously described[39]. In brief, a Bio-Bond 5U C4 column (Dikma) was used to achieve separation of lipids. The liquid chromatography solvents were as follows: buffer A, 95:5 H_2O:methanol + 0.03% NH_4OH; buffer B, 60:35:5 isopropanol:methanol:H_2O + 0.03% NH_4OH. A typical liquid chromatography run consisted of the following for 70 min after injection: 0.1 ml min^{-1} 100% buffer A for 5 min, 0.4 ml min^{-1} linear gradient from 20% buffer B to 100% buffer B over 50 min, 0.5 ml min^{-1} 100% buffer B for 8 min and equilibration with 0.5 ml min^{-1} 100% buffer A for 7 min. Lipidomic analysis was performed using a Thermo Fisher Scientific Q Exactive Plus fitted with a heated electrospray ionization source in negative ionization mode. The MS source parameters were 4-kV spray voltage, with a probe temperature of 437.5 °C and capillary temperature of 268.75 °C. Full-scan mass spectrometry data was collected at a resolution of 70 k, automatic gain control target 1 × 106, maximum injection time of 100 ms and a scan range of 150–2,000 *m*/*z*. Data-dependent mass spectrometry (top 5 mode) was acquired at a resolution of 35 k, automatic gain control target 1 × 105, maximum injection time of 50 ms, isolation window 1 *m*/*z*, scan range 200–2,000 *m*/*z* and a stepped normalized collision energy of 20, 30 and 40. Extracted ion chromatograms for each lipid were generated using a threshold of 5 p.p.m. around the molecular anion $[\mathrm{M-H}]^-$ exact mass. Lipids, acyl chain composition and degree of unsaturation were validated by analysing the product ions in the corresponding tandem mass spectra. Relative quantification of lipids was performed by measuring the area under the peak and dividing this number by the area under the peak for the internal standard $^{13}\mathrm{C}_{16}$-palmitatic acid.

***In vivo* experiments.** Mice were purchased from The Jackson Laboratories. All the colonies were bred as indicated by Jackson Laboratories and maintained in pathogen-free conditions at The Salk Institute, except NOD.Cg-*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/SzJ mice, which were maintained at The University of Texas MD Anderson Cancer Center and *Tyr-Nras*^{Q61K} mice and their wild-type counterparts, which were maintained at the University of California, Irvine. C57BL/6J, NOD.Cg-*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/SzJ and *Tyr-Nras*^{Q61K} 3–14-week-old male and female mice were used. No statistical methods were used to predetermine sample size but it was determined according to previous experimental observations. All the procedures performed in this study were approved by the Institutional Animal Care and Use Committee of the SALK institute, the University of California, Irvine and the University of Texas MD Anderson Cancer Center. In all experiments, mice were monitored daily for signs of illness and were euthanized when they reached endpoints. The 005 cells (5×10^4 cells per 1.5 μl) or GSC 8.11 cells (1.5×10^5 per 1.5 μl) were stereotaxically injected into anaesthetized 8–16-week-old mice (C57BL/6J for 005 cells, and NOD.Cg-*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/SzJ for GSCs). The following coordinates were used: subventricular zone, 1.5 mm, 2.0 mm and 2.3 mm; hard palate, 2.0 mm, 1.5 mm and 2.3 mm; cerebral cortex, 1 mm, 1 mm, and 0.5 mm or 1.0 mm; 0 mm, 1 mm and 0.5 mm or 1.0 mm; and 2.0 mm, 1.5 mm and 0.5 mm; striatum, 0 mm, 1.4 mm and 3.0 mm (all measurements are posterior, lateral and dorsal to the bregma, respectively). To ensure that each experimental group had an equivalent starting tumour, after tumour sizes were quantified by magnetic resonance imaging mice were divided into two groups (vehicle and SR9009) and three weeks after injection, the treatment was started. Similarly, NOD.Cg-*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/SzJ mice were imaged by bioluminescence imaging and after tumour sizes were quantified they were divided in two (vehicle and SR9009) or three groups (vehicle, SR9009 and temozolomide). The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment. For all bioluminescence imaging, D-luciferin (150 mg kg^{-1}) was administered by subcutaneous injection to mice 10 min before imaging. Mice were fed with Picolab Diet 20 No. 5058. SR9009 was administered twice per day by intraperitoneal injection at the indicated concentrations for one week, and on subsequent days once a day unless otherwise stated. Temozolomide was administered once per day by intraperitoneal injection at 82.5 mg kg^{-1} for 5 days. All mice in this study were kept according to guidelines approved by the Animal Care and Use Committee of the Salk Institute. For the naevi studies, both ears of *Tyr-Nras*^{Q61K} and wild-type littermate mice at postnatal day 21 were treated with SR9009 or DMSO. Forty microlitres of drug SR9009 at 20 μM diluted with DMSO was applied to each ear twice daily for twelve consecutive days. Mice were killed one hour after the final treatment. Four mice (eight ears) were used in each group.

Statistical analysis. Results are shown as means ± s.d. or s.e.m., as indicated in the figure legends. *P* values were calculated as indicated in figure legends, with 95% confidence level.

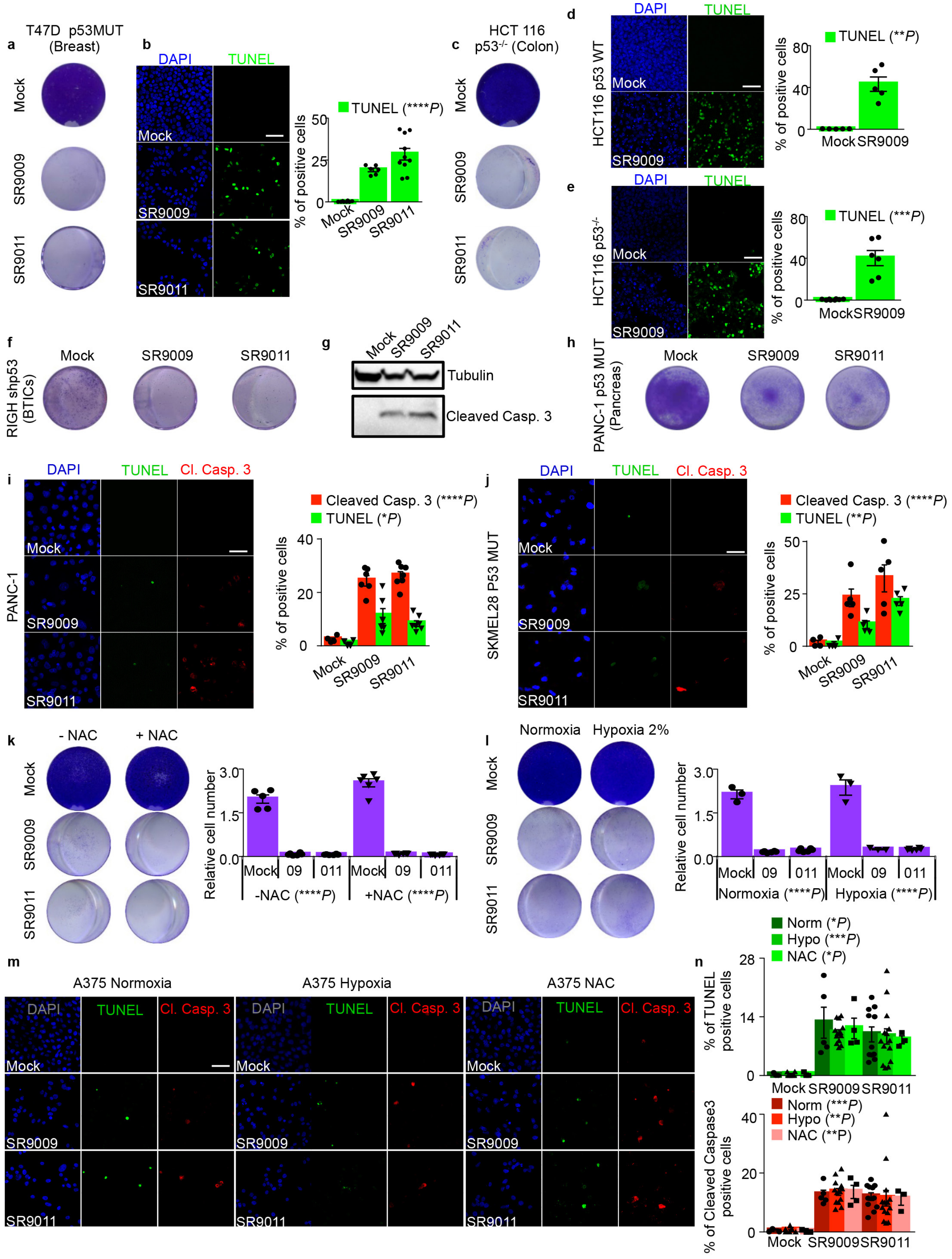
Data availability. All data and reagents are available from the authors upon reasonable request. All gel source data are available in Supplementary Fig. 1. Tumour source data (Fig. 4f and Extended Data Fig. 9d, i) are available as Source Data; ChIP–seq data[5] and The Cancer Genome Atlas data[30,31] have previously been published.

32. Marumoto, T. *et al.* Development of a novel mouse glioma model using lentiviral vectors. *Nat. Med.* **15,** 110–116 (2009).
33. Di Micco, R. *et al.* Interplay between oncogene-induced DNA damage response and heterochromatin in senescence and cancer. *Nat. Cell Biol.* **13,** 292–302 (2011).
34. Hossain, A. *et al.* Mesenchymal stem cells isolated from human gliomas increase proliferation and maintain stemness of glioma stem cells through the IL-6/gp130/STAT3 pathway. *Stem Cells* **33,** 2400–2415 (2015).
35. Zhang, Y. *et al.* Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9,** R137 (2008).
36. Lam, M. T. *et al.* Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* **498,** 511–515 (2013).
37. Bligh, E. G. & Dyer, W. J. A rapid method of total lipid extraction and purification. *Can. J. Biochem. Physiol.* **37,** 911–917 (1959).
38. Saghatelian, A. *et al.* Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry* **43,** 14332–14339 (2004).
39. Svensson, R. U. *et al.* Inhibition of acetyl-CoA carboxylase suppresses fatty acid synthesis and tumor growth of non-small-cell lung cancer in preclinical models. *Nat. Med.* **22,** 1108–1119 (2016).



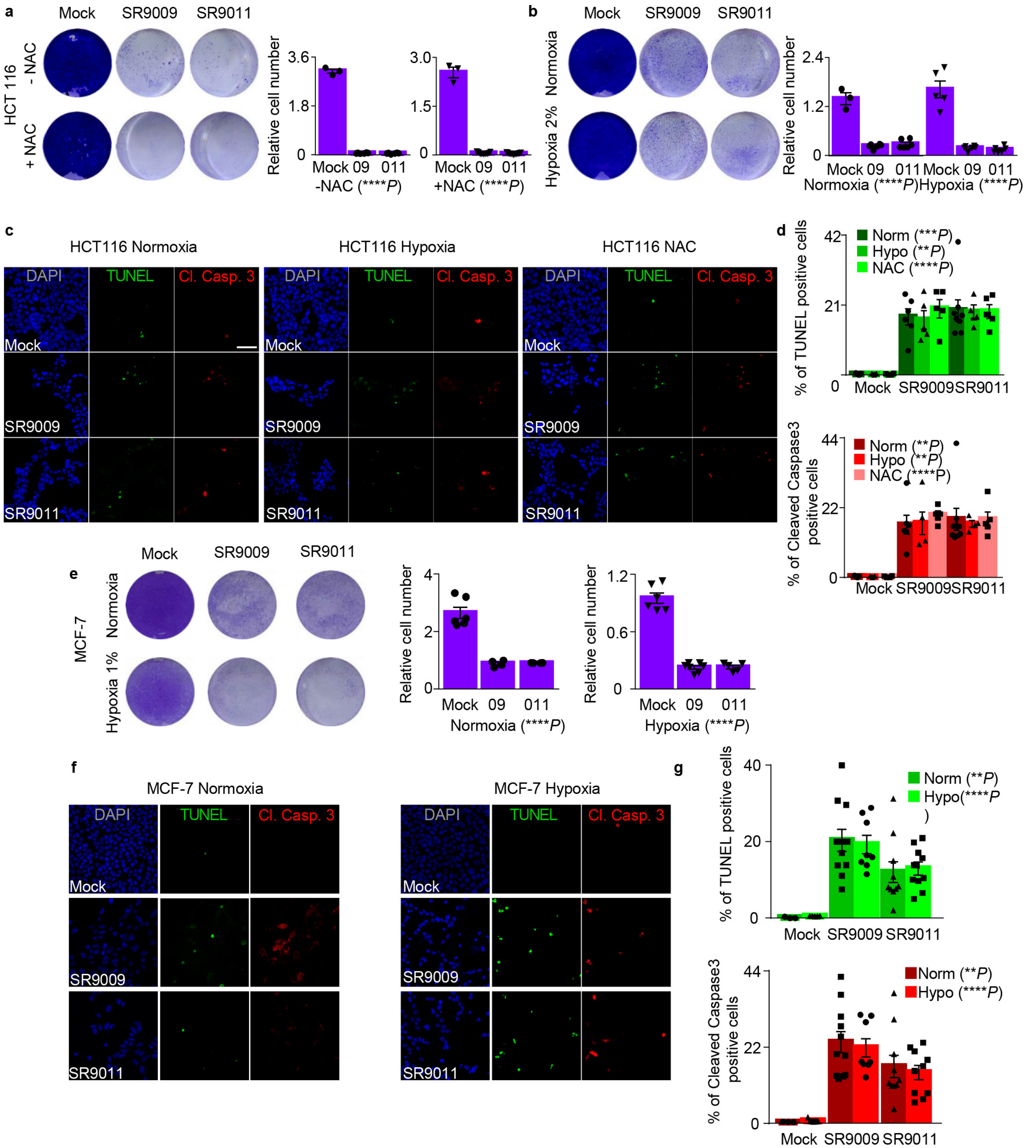
Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | SR9011, an additional agonist of REV-ERBs, selectively kills cancer cell lines. a, Viability assay shows that SR9011 is cytotoxic specifically in cancer cells (72 h). One-way ANOVA. *n* indicates biological replicates: astrocytes, $n = 7$ (mock), $n = 7$ (2.5 μM), $n = 9$ (5 μM), $n = 13$ (10 μM) and $n = 13$ (20 μM), $^{***}P = 0.0004$; astrocytomas, $n = 21$ (mock), $n = 15$ (2.5 μM), $n = 7$ (5 μM), $n = 8$, (10 μM), $n = 7$ (20 μM), $^{****}P < 0.0001$; BTICs, $n = 10$ (mock), $n = 8$ (2.5 μM), $n = 9$ (5 μM), $n = 13$ (10 μM), $n = 10$ (20 μM), $^{****}P < 0.0001$. **b**–**d**, Proliferation assay showing that SR9011 treatment does not affect normal BJ cells, but is deleterious for transformed BJ-ELR cells and the cancer cell lines MCF-7 and HCT116 (20 μM, 7 days). Depletion of REV-ERBs by shRNA impairs apoptosis induction by the SR9011 agonist of REV-ERBs; $n = 3$ biologically independent experiments. **e**, Human acute T-cell leukaemia cells are affected by the SR9011 agonist of REV-ERBs (72 h, one-tailed Mann–Whitney test, $^{****}P < 0.0001$; $n = 24$ (mock) and 12 (SR9011) biological replicates). **f**, Immunoblot analysis of cleaved caspase 3 shows that agonists of REV-ERBs trigger apoptosis in the A375 melanoma cell line (representative of $n = 2$ biologically independent experiments). **g**–**j**, Immunostaining for cleaved caspase 3 and TUNEL assay confirm apoptosis induction by SR9011 in the cancer cell lines MCF-7 and A375. **h**, **j**, quantification of **g** and **i**, respectively. *n* indicates biologically independent samples; MCF-7, $n = 5$ (mock) or 11 (SR9011); A375, $n = 8$ (mock) or 16 (SR9011). One-tailed Mann–Whitney test, MCF-7 cleaved caspase 3, $^{*}P = 0.0117$; TUNEL assay, $^{*}P = 0.0231$; A375 cleaved caspase 3, $^{****}P < 0.0001$; TUNEL assay, $^{****}P < 0.0001$. Scale bars, 50 μm. **k**, Electron microscopy confirms induction of apoptosis, as indicated by extensive presence of swollen mitochondria (representative of $n = 3$ biologically independent samples in two experiments). Arrows, normal mitochondria; asterisks, swollen mitochondria. Nu, nucleus. Scale bar, 1 μm. **l**, Downregulation of *NR1D1* and *NR1D2* is confirmed by qRT–PCR (A375). $n = 3$ biologically independent samples. One-tailed Mann–Whitney test, $^{*}P = 0.05$. All panels representative of three biologically independent experiments unless otherwise specified. All data are mean ± s.e.m. a.u., arbitrary units. For gel source data, see Supplementary Fig. 1.

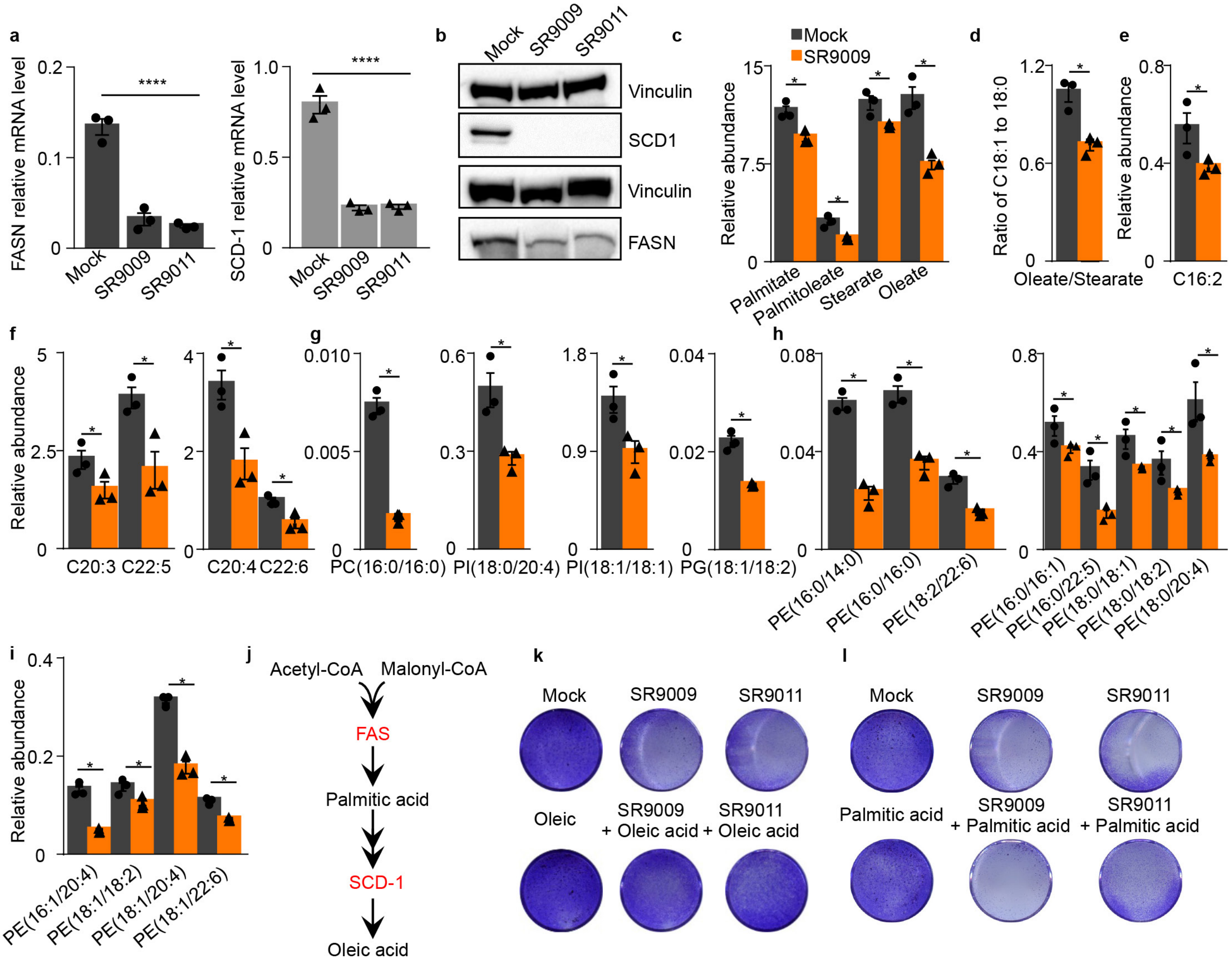


Extended Data Figure 2 | See next page for caption.

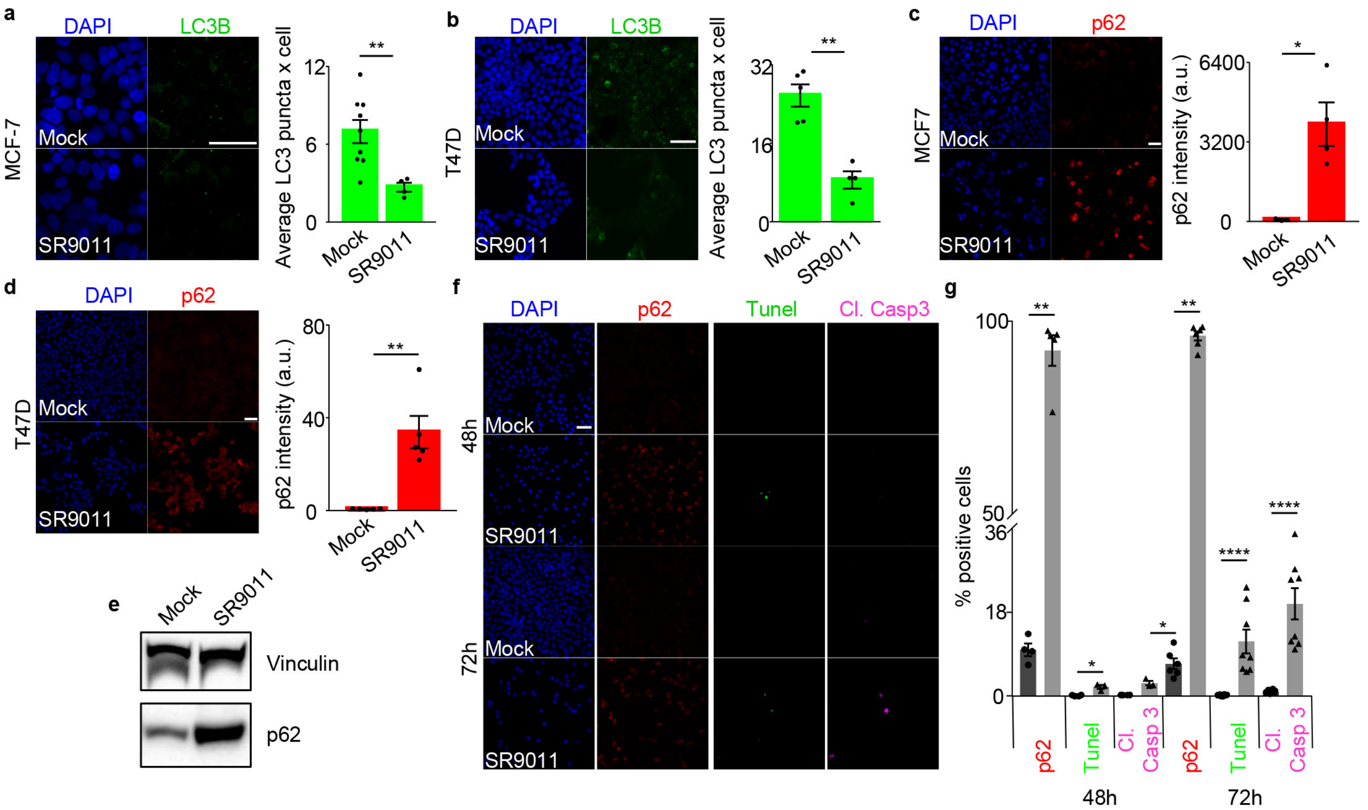
Extended Data Figure 2 | Induction of apoptosis by agonists of REV-ERBs is independent of p53 and oxidative stress. a–**j**, Treatment with agonists of REV-ERBs triggers apoptosis independently of p53 status, as shown by proliferation assay (7 days, 20 μM; **a**, **c**, **f**, **h**) and TUNEL assay (3 days, 20 μM; **b**, **i**, **j**) in cancer cell lines affected by various types of p53 alteration. *n* indicates biologically independent samples; T47D, $n = 8$ (mock), $n = 6$ (SR9009) and $n = 10$ (SR9011), one-way ANOVA, $****P < 0.0001$; PANC-1, $n = 4$ (mock), $n = 6$ (SR9009) and $n = 7$ (SR9011), one-way ANOVA; TUNEL assay, $*P = 0.0108$; cleaved caspase 3, $****P < 0.0001$; SKMEL28, $n = 4$ (mock), $n = 5$ (SR9009, SR9011), one-way ANOVA; TUNEL assay, $****P < 0.0001$; cleaved caspase 3, $**P = 0.0035$. **d**, **e**, Apoptosis is induced in both wild-type and p53-null HCT116 cells (TUNEL assay, 4 days, 20 μM, mean ± s.e.m.). One-tailed Mann–Whitney test. *n* indicates biologically independent samples. HCT116 wild type, $n = 5$ (mock, SR9009), $**P = 0.004$; HCT116 p53 knockout, $n = 8$ (mock) or 6 (SR9009), $***P = 0.0003$. **f**, Immunoblot analysis of cleaved caspase 3 shows that agonists of REV-ERBs trigger apoptosis in the RIGH cell line (one experiment). **k**, Co-treatment with the reducing agent NAC (10 mM) does not rescue the viability of A375 cells (20 μM, 7 days). *n* indicates biological replicates, $n = 5$ (mock, −NAC) or 6 (all other dot plots). One-way ANOVA, $****P < 0.0001$. **l**, Results obtained under hypoxic conditions (20 μM, 6 days) were similar to those obtained in co-treatments with NAC. *n* indicates biological replicates, $n = 3$ (mock −NAC, mock hypoxia and SR9009 hypoxia) or 6 (SR9009 normoxia, SR9011 normoxia and SR9011 hypoxia). One-way ANOVA, $****P < 0.0001$). **m**, **n**, Hypoxia or co-treatment with NAC does not alter the ability of agonists of REV-ERBs to induce apoptosis in A375 cells. One-way ANOVA. *n* indicates biologically independent samples: normoxia, $n = 3$ (mock), $n = 5$ (SR9009) or $n = 11$ (SR9011); TUNEL assay, $*P = 0.0432$; cleaved caspase 3, $***P = 0.0004$; hypoxia, $n = 6$ (mock), $n = 13$ (SR9009) or $n = 14$ (SR9011); TUNEL assay, $***P = 0.0005$; cleaved caspase 3, $*P = 0.0028$; NAC, $n = 3$ (mock), $n = 4$ (SR9009) or $n = 3$ (SR9011); TUNEL assay, $*P = 0.0104$; cleaved caspase 3, $**P = 0.0042$. Scale bars, 50 μm. All panels representative of three biologically independent experiments with similar results unless otherwise specified. All data are mean ± s.e.m. Norm, normoxia; Hypo, hypoxia. For gel source data, see Supplementary Fig. 1.



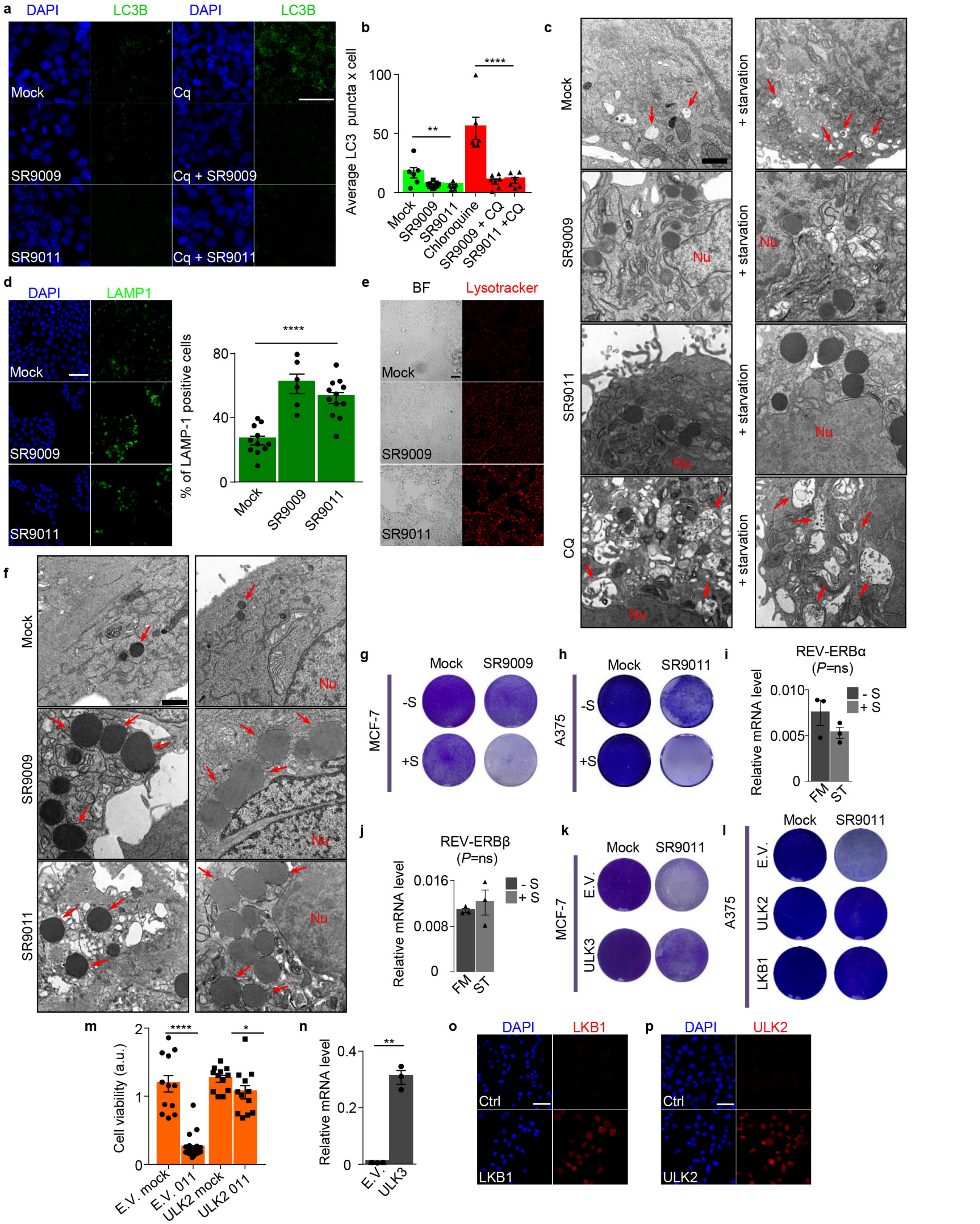
Extended Data Figure 3 | Attenuation of oxidative stress does not affect the cytotoxic activity of agonists of REV-ERBs. a, b, Treatment with agonists of REV-ERBs (20 μM) induces apoptosis on co-treatment with NAC and under hypoxic conditions, as shown by proliferation assays of HCT-116 cells. *n* indicates biological replicates. *n* = 3 (mock ± NAC), *n* = 6 (SR9009/SR9011 ± NAC), *n* = 9 (SR9009 + NAC), *n* = 11 (SR9011 + NAC), *n* = 3 (mock normoxia), *n* = 6 (SR9009/SR9011 normoxia and hypoxia), *n* = 5 (mock hypoxia) 6 days, one-way ANOVA, ****$P < 0.0001$. **c, d**, Apoptosis induction of HCT116 cells remained unchanged under hypoxia or on co-treatment with NAC; 20 μM, 6 days, one-way ANOVA. *n* indicates biologically independent samples: normoxia, *n* = 5 (mock), *n* = 6 (SR9009), *n* = 8 (SR9011); TUNEL assay, ***$P = 0.0003$; cleaved caspase 3, **$P = 0.0021$; hypoxia, *n* = 3 (mock), *n* = 5 (SR9009), *n* = 4 (SR9011); TUNEL assay, **$P = 0.0015$; cleaved caspase 3, **$P = 0.0046$; NAC, *n* = 4 (mock), *n* = 5 (SR9009), *n* = 5 (SR9011); TUNEL assay, ****$P < 0.0001$; cleaved caspase 3, ****$P < 0.0001$. **e**, In MCF-7 cells, apoptosis triggered by agonists of REV-ERBs is independent of oxidative state, as shown by proliferation assay (20 μM). *n* indicates biological replicates: *n* = 6 (mock, normoxia and hypoxia), *n* = 4 (09-011 normoxia), *n* = 7 (09 hypoxia), *n* = 5 (011 hypoxia); one-way ANOVA, ****$P < 0.0001$; **f, g** TUNEL assay and immunofluorescence analysis of cleaved caspase 3 confirm previous results (Extended Data Fig 3c, d). *n* indicates biologically independent samples. *n* = 3 (mock normoxia), *n* = 5 (mock hypoxia), *n* = 11 (09 normoxia), *n* = 8 (09 hypoxia), *n* = 10 (011 normoxia and hypoxia). One-way ANOVA. Normoxia, TUNEL assay, **$P = 0.0049$; cleaved caspase 3,**$P = 0.0054$; hypoxia, TUNEL assay, ****$P < 0.0001$; cleaved caspase 3, ****$P < 0.0001$. All panels representative of three biologically independent experiments with similar results. All data are mean ± s.e.m. Norm, normoxia; Hypo, hypoxia.



Extended Data Figure 4 | Agonists of REV-ERBs inhibit *de novo* lipogenesis. a, **b**, Agonists of REV-ERBs downregulate *FASN* and *SCD1* mRNA, as assayed by qRT–PCR. A172 cell line 48 h, 20 μM; FASN and SCD1, $n = 3$ biologically independent samples, $****P < 0.0001$. FASN and SCD1 protein levels (**b**) are reduced on treatment with SR9009 and SR9011. **c**–**i**, Agonists of REV-ERBs reduce free fatty acid concentrations, as quantified by liquid chromatography–mass spectrometry. **c**, Relative levels of free fatty acids that are the primary products of FASN (palmitic acid, C16:0; stearic acid, C18:0) and SCD1 (palmitoleic acid, C16:1; oleic acid, C18:1) are lower in samples treated with SR9009 than in control samples. **d**, The unsaturation index (changes in oleate:stearate ratio) is decreased in the SR9009-treated sample, compared to mock-treated samples, owing to the large decrease in monounsaturated oleate observed in the SR9009-treated sample. **e**, **f**, SR9009 treatment reduces polyunsaturated fatty acids levels compared to mock-treated samples, in agreement with previous results (Extended Data Fig. 3c, d). **g**–**i**, Decreases in free fatty acid levels can affect the concentrations of phospholipids that contain these fatty acids. Treatment with agonists of REV-ERBs leads to reductions in palmitic acid-containing phosphatidylcholine, arachidonic acid- and oleic acid-containing phosphatidylinositols, mono- and di-unsaturated phosphatidylglycerol (**g**) and phosphatidylethanolamines (**h**, **i**); A172 cell line 48 h, 20 μM, $*P = 0.05$. **j**, Simplified scheme illustrating the metabolic products of FASN and SCD1. **k**, Supplementation of oleic acid partially ameliorates the cytotoxity of agonists of REV-ERBs (A172, 20 μM, 4 days). **l**, Supplementation of palmitic acid does not impair the cytotoxicity of agonists of REV-ERBs (20 μM, A172, 4 days). All data are mean ± s.e.m. *P* value is calculated with one-way ANOVA in panel **a**, and with one-tailed Mann–Whitney test in the remaining panels. $n = 3$ biologically independent samples (**d**–**i**). For gel source data, see Supplementary Fig. 1.

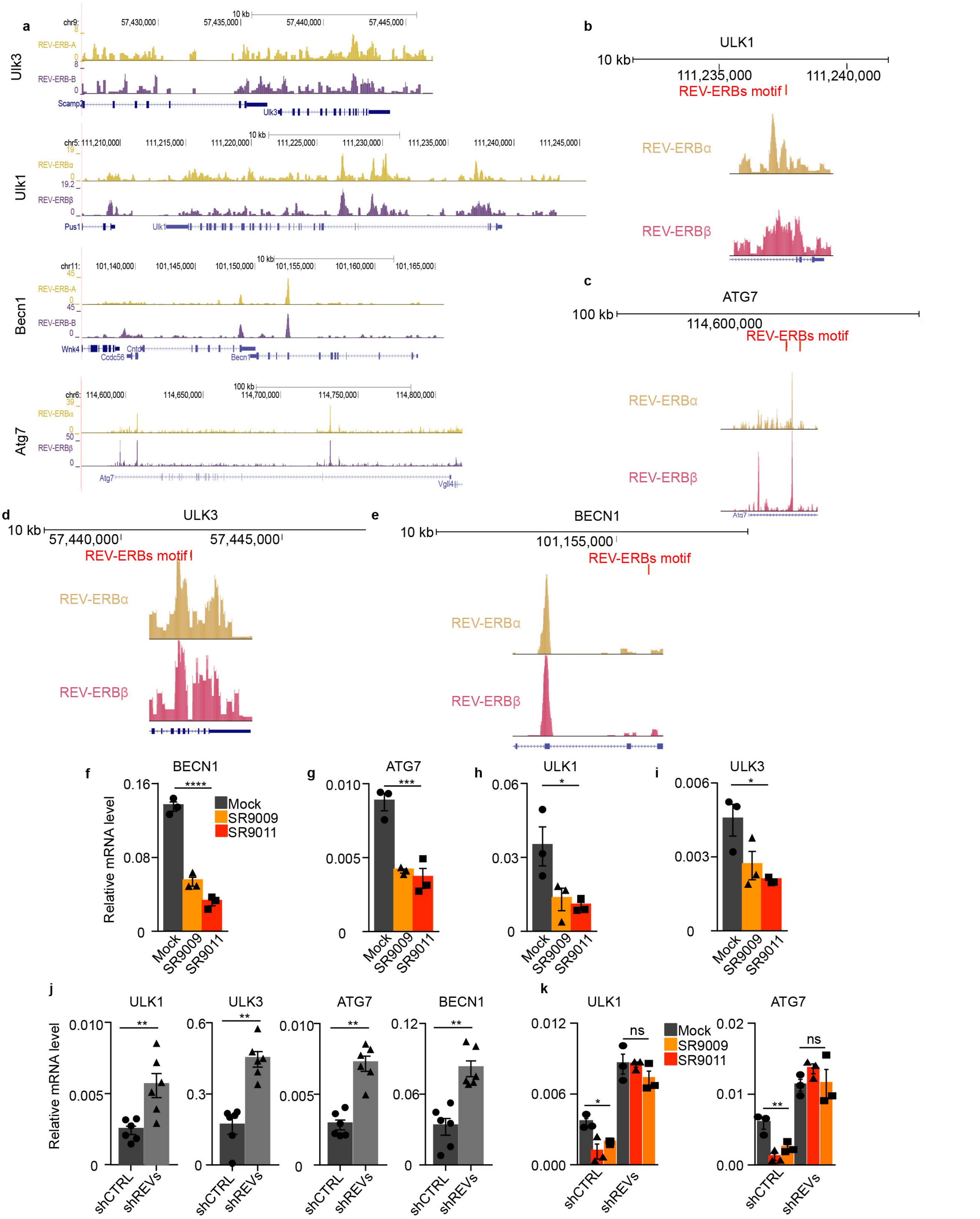


Extended Data Figure 5 | REV-ERB agonist SR9011 inhibit autophagy. **a**, **b**, Treatment with SR90011 reduces the number of autophagosomes both in MCF7 and T47D cell lines. *n* indicates biologically independent samples, MCF7 20 μM 24 h, $n = 9$ (mock), $n = 4$ (SR9011), $^{**}P = 0.0056$; T47D 48 h 20 μM, $n = 5$ (mock), $n = 4$ (SR9011), $^{**}P = 0.0079$. **c**, **d**, SR9011 induces accumulation of p62, as shown by immunofluorescence both in MCF7 and T47D cell lines. *n* indicates biologically independent samples. MCF7 p62 48 h, $n = 3$ (mock), $n = 4$ (SR9011), $^{*}P = 0.0286$; T47D 48 h, $n = 5$ (mock and SR9011), $^{**}P = 0.004$. **e**, Accumulation of p62 is confirmed by immunoblot (48h, 20 μM A375). **f**, **g**, Inhibition of autophagy precedes apoptosis induction, as shown by immunofluorescence of p62, cleaved caspase 3 and TUNEL assay. *n* indicates biologically independent samples: $n = 4$ (mock, 48 h), $n = 5$ (SR9011 p62), $n = 3$ (SR9011, 48 h), $n = 6$ (mock, SR9011 72 h, p62), $n = 10$ (mock, 72 h), $n = 8$ (SR9011, 72 h) A375 20 μM; cleaved caspase 3, 48 h, $^{*}P = 0.0286$; cleaved caspase 3, 72 h, $^{****}P < 0.0001$; TUNEL assay, 48 h, $^{*}P = 0.0286$; TUNEL assay, 72 h, $^{****}P < 0.0001$; p62, 48 h, $^{**}P = 0.0079$; p62, 72 h, $^{**}P = 0.0011$. All panels representative of three biologically independent experiments with similar results. All data are mean ± s.e.m. *P* value is calculated with one-tailed Mann–Whitney test in all panels. For gel source data, see Supplementary Fig. 1.

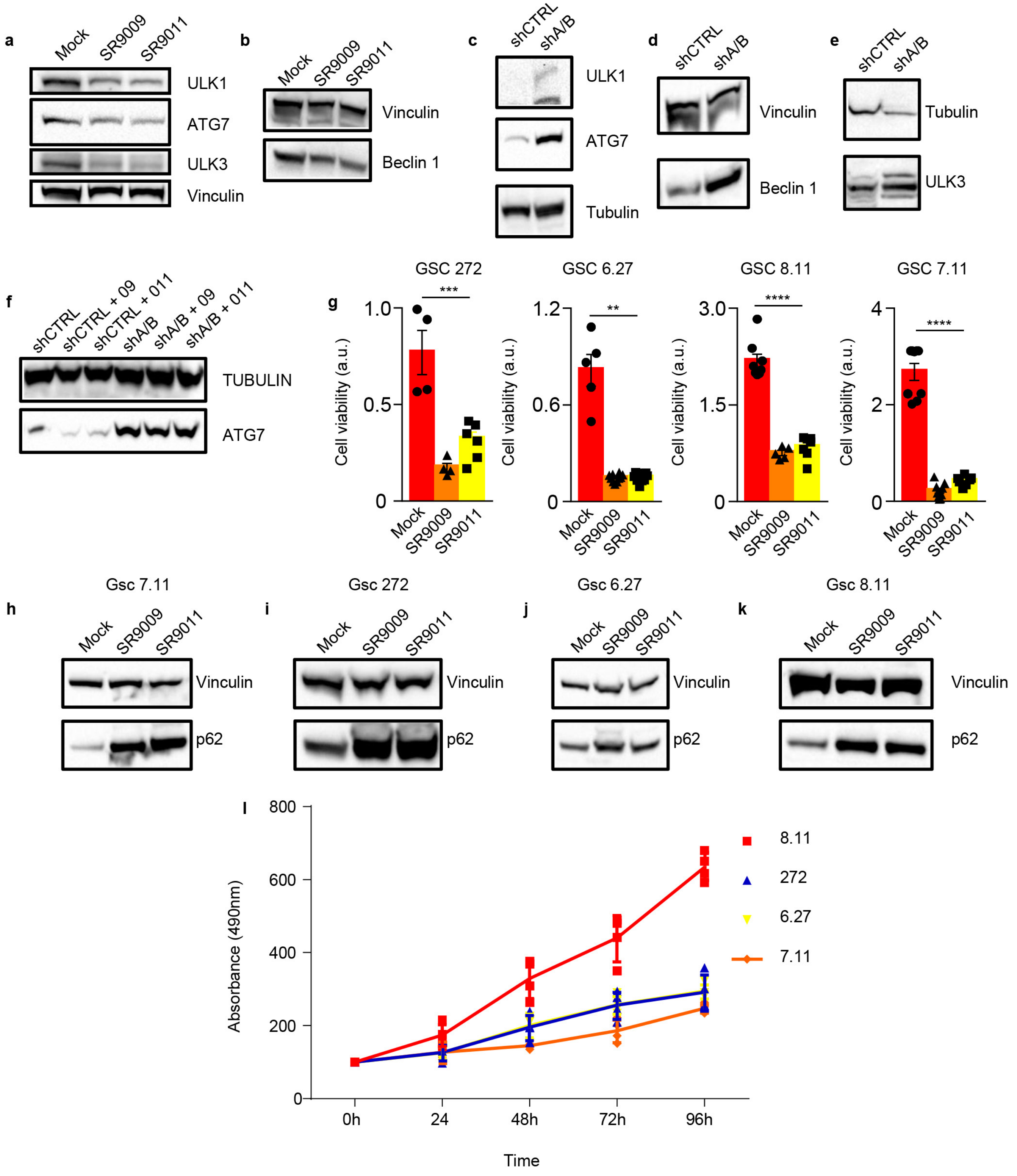


Extended Data Figure 6 | See next page for caption.

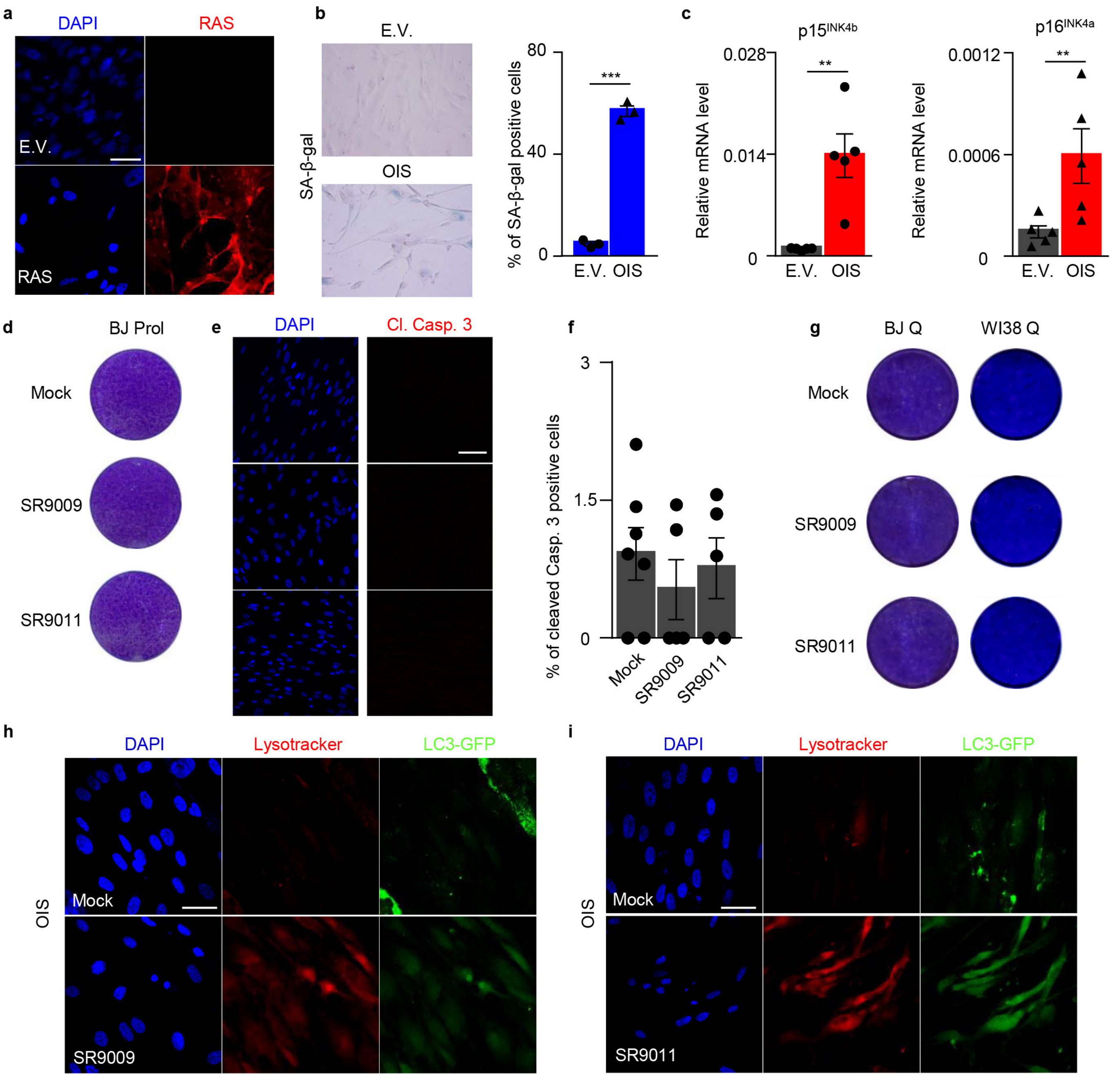
Extended Data Figure 6 | Agonists of REV-ERBs (SR9009 and SR9011) block autophagy. a, Agonists of REV-ERBs block autophagy, which results in reduced autophagic flux. **b**, Quantification of LC3 puncta. *n* indicates biologically independent samples. $n = 6$ (mock, chloroquinine (CQ) ± SR9011), $n = 11$ (SR9009), $n = 5$ (SR9011), $n = 7$ (CQ + SR9009). One-way ANOVA: mock versus SR9009 and SR9011, $^{**}P = 0.0049$; CQ versus CQ + SR9009 or CQ + SR9011, $^{****}P < 0.0001$. **c**, On treatment with SR9009 and SR9011, autophagy blockage can be observed by electron microscopy, even under starvation conditions. Arrows, representative autophagosomes; Nu, nucleus. Scale bars, 1 μM. $n = 3$ biologically independent samples of two independent experiments with similar results (mock ± SR9009 and SR9011) or one experiment (mock, SR9009 and SR9011 ± starvation). **d**, Agonists of REV-ERBs induce lysosome accumulation, as shown by immunofluorescence assay for the lysosome marker LAMP1. *n* indicates biologically independent samples: $n = 11$ (mock), $n = 6$ (SR9009), $n = 12$ (SR9011). T47D, 72 h 20 μM, one-way ANOVA, $^{****}P < 0.0001$. **e**, Lysosome accumulation is confirmed by LysoTracker Red (MCF-7, 72 h 20 μM). Scale bars, 50 μm. BF, bright field. **f**, Marked lysosomal turnover defects are revealed with electron microscopy. $n = 3$ biologically independent samples of two independent experiments with similar results. Arrows, lysosomes; Nu, nucleus. Scale bars, 1 μM. **g**, **h**, Starvation synergizes with treatment with the SR9009 agonist of REV-ERBs (MCF-7 48 h, 20 μM; A375, 3 days, 20 μM). **i**, **j**, Starvation has no effect on expression of REV-ERBs, as shown by qRT–PCR; two-tailed Mann–Whitney test. ns, not significant; FM, fresh medium; ST, starvation. **k**, **l**, Overexpression of ULK3, ULK2 and LKB1 impairs the induction of apoptosis by SR9011 (MCF-7, A375 6 days, 20 μM). **m**, WST-1 viability assay shows abrogation of apoptosis in cells that overexpress ULK2. *n* indicates biological replicates: $n = 12$ (empty vector (E.V.) mock, ULK2 mock, ULK2 SR9011), $n = 27$ (E.V. SR9011) A375, 6 days, 20 μM. One-tailed Mann–Whitney test, E.V. mock versus E.V. 011, $^{****}P < 0.0001$; ULK2 mock versus ULK2 011 $^{*}P = 0.0225$). **n**, qRT–PCR shows overexpression of ULK3 (one-tailed Student's *t*-test, $^{**}P = 0.0031$). **o**, **p**, Immunofluorescence assay confirms overexpression of LKB1 and ULK2. $n = 3$ biologically independent samples (**i**, **j**, **n**). Scale bars, 50 μm. All panels representative of three biologically independent experiments with similar results, unless otherwise specified. All data are mean ± s.e.m.



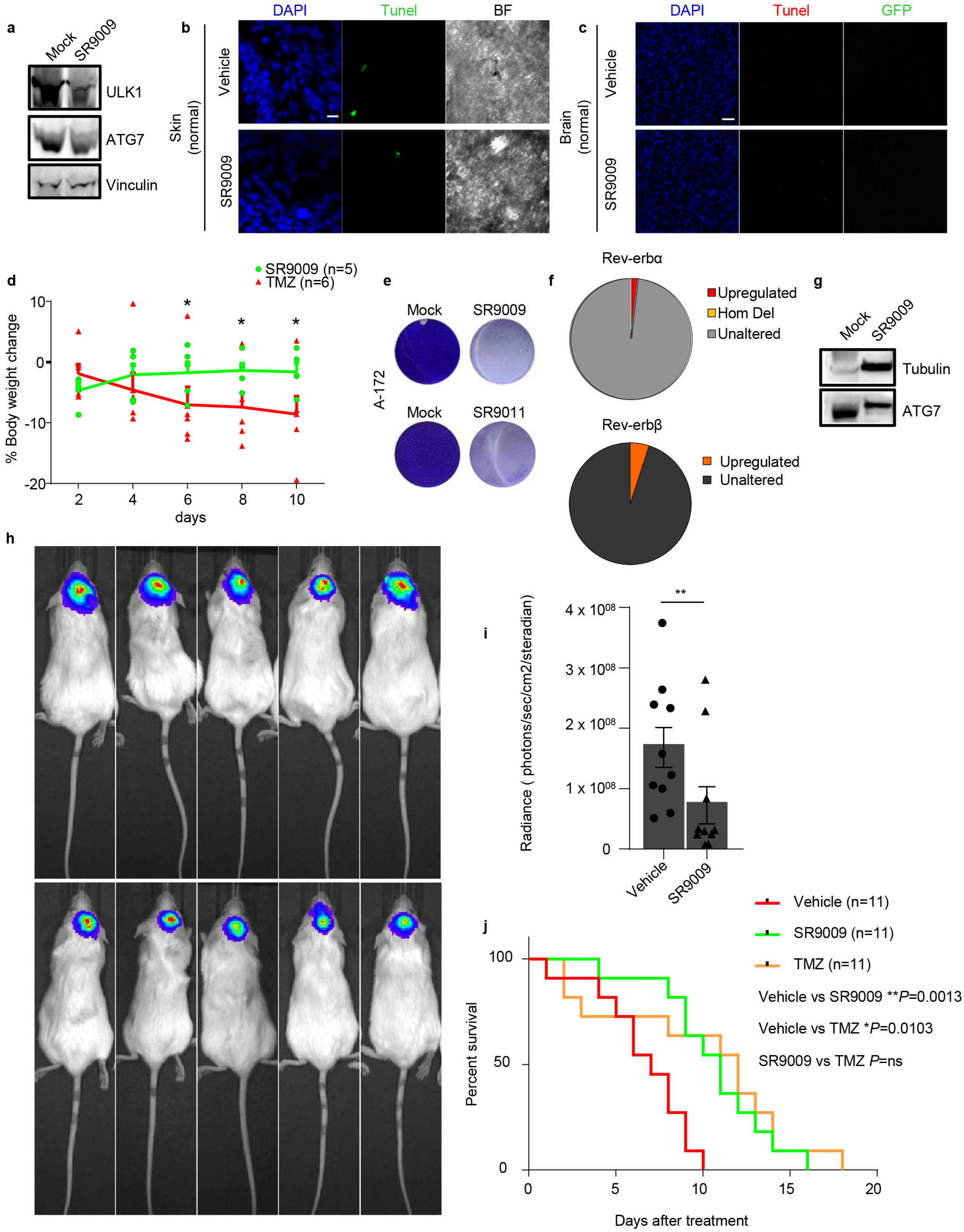
Extended Data Figure 7 | Core autophagy genes are novel REV-ERBs targets. a, Analyses of available ChIP–seq data[5] indicate that REV-ERBs peaks are present in *Ulk3*, *Ulk1*, *Becn1* and *Atg7* ($P < 0.00001$ calculated by MACS using Poisson distribution, false discovery rate ≤ 0.05). **b–e**, Analysis of REV-ERBs-binding motif performed using HOMER indicates the presence of REV-ERBs-binding sites in *Ulk3*, *Ulk1*, *Becn1* and *Atg7* genes (mouse genome). **f–i**, Treatment with agonists of REV-ERBs leads to downregulation of autophagy central regulators (MCF-7 72 h 20 μM; one-way ANOVA, $^{****}P < 0.0001$). **j**, Autophagy genes are upregulated on expression of REV-ERBs shRNA. A375 cell line, $n = 6$ biologically independent samples. One-tailed Mann–Whitney test; *ULK3*, $^{**}P = 0.0011$; *ATG7* and *BECN1*, $^{**}P = 0.0011$; *ULK1*, $^{**}P = 0.0043$. **k**, The repression of autophagy genes caused by SR9009 and SR9011 is abrogated in A375 cells expressing REV-ERBs shRNA; control shRNA ± SR9009 or SR9011, one-way ANOVA; ULK1, $^{*}P = 0.0162$; ATG7, $^{**}P = 0.0036$. **f–i**, **k**, $n = 3$ biologically independent samples. All data are mean ± s.e.m. shREVs, shNR1D1 and shNR1D2; shCTRL, non-silencing shRNA.



Extended Data Figure 8 | REV-ERBs regulate autophagy core genes and block autophagy in slowly proliferating cancer stem cells. **a**, **b**, Immunoblot analyses show a reduction of ULK1, ATG7, ULK3 and BECN1 protein levels on treatment with agonists of REV-ERBs (MCF-7 72 h 20 μM). **c**–**e**, REV-ERBs shRNA increases protein levels of autophagy regulators (A375). **f**, The reduction in ATG7 protein levels on treatment with SR9009 and SR9011 is abrogated in cells expressing REV-ERBs shRNA. **g**, WST-1 viability assays show that treatment with SR9009 and SR9011 is cytotoxic specifically in patient-derived glioblastoma stem cells; mean ± s.e.m., 5 days, one-way ANOVA. *n* indicates biological replicates: GSC 272, $n = 4$ (mock, SR9009), $n = 6$ (SR9011), $^{***}P = 0.0002$; GSC 6.27, $n = 5$ (mock), $n = 10$ (SR9009), $^{**}P = 0.003$; GSC 8.11, $n = 8$ (mock), $n = 5$ (SR9009), $n = 7$ (SR9011), $^{****}P < 0.0001$; GSC 7.11, $n = 9$ (mock, SR9009), $n = 7$ (SR9011), $^{****}P < 0.0001$. **h**–**k**, Immunoblot analyses show accumulation of p62 in patient-derived glioblastoma stem cells (one independent experiment). **l**, MTS assays show that GSC 6.27, 7.11 and 272 are characterized by a slow proliferation rate; $n = 4$ biologically independent samples, four experiments, mean ± s.d. All panels representative of three biologically independent experiments with similar results, unless otherwise specified. For gel source data, see Supplementary Fig. 1.



Extended Data Figure 9 | Agonists of REV-ERBs do not affect viability of normal proliferating and quiescent OIS cells. a, Immunofluorescence assay for RAS confirms RAS overexpression in OIS cells. **b**, SA-β-galactosidase assay shows induction of senescence; $n = 3$ biologically independent samples, one-tailed Student's t-test, $****P < 0.0001$. **c**, Induction of cell cycle inhibitors *CDKN2B* and *CDKN2A* is assayed by qRT–PCR. $n = 5$ biologically independent samples. One-tailed Mann–Whitney test, *CDKN2B*, $**P = 0.004$; *CDKN2A*, $**P = 0.0079$. **d**, **e**, Agonists of REV-ERBs do not induce apoptosis in proliferating and quiescent normal diploid BJ fibroblasts (**d**–**g**), as shown by proliferation assay (**d**, 7 days, 20 μM) and immunofluorescence for cleaved caspase 3 (**e**, **f**, 7 days, 20 μM). One-way ANOVA. n indicates biologically independent samples: $n = 7$ (mock), $n = 5$ (SR9009, SR9011). Cell viability is also not affected in an additional normal diploid WI38 cell line (**g**, 10 days, 20 μM). **h**, **i**, SR9009 and SR9011 inhibit autophagy in OIS cells, as shown by the accumulation of lysosomes (LysoTracker Red) and the absence of LC3 puncta (3 days, 20 μM). Scale bars, 50 μm. Data in **a**–**i** are representative of three independent experiments with similar results, unless otherwise specified. All data are mean ± s.e.m.



Extended Data Figure 10 | See next page for caption.

Extended Data Figure 10 | SR9009 impairs tumour growth and improves survival of glioblastoma patient-derived xenografts. a, Protein levels of autophagy genes in NRAS naevi are reduced upon treatment with SR9009, as assayed by immunoblot (*n* = 4 mice, one experiment). **b**, TUNEL assays show that apoptosis induction is absent in normal skin on treatment with SR9009 (*n* = 4 mice, one experiment, 12 days, SR9009 20 μM). Scale bar, 10 μm; BF, bright field. **c**, TUNEL assays show that apoptosis induction is absent in normal brain tissues on treatment with SR9009 (6 days, 200 mg kg^{-1} twice a day, *n* = 5 mice, one experiment). **d**, Treatment with SR9009 (100 mg kg^{-1} twice a day) is tolerated better than temozolomide administration (82.5 mg kg^{-1} once a day for 5 days), as shown by measurement of percentage body weight change. One-tailed Mann–Whitney test; day 6, 8, 10, **P* = 0.0411, mean ± s.e.m. *n* = 5 (SR9009) or 6 (temozolomide) mice. **e**, The glioblastoma cell line A172 is sensitive to treatment with SR9009 and SR9011 (20 μM 6 days, three biologically independent experiments with similar results). **f**, Previous analyses[30] of The Cancer Genome Atlas data show genetic alterations that affect *NR1D1* and *NR1D2* are absent. Gene expression analysis shows that no cases are present with downregulation of REV-ERBs, and only a small fraction with upregulation. *n* = 574 biologically independent samples. *NR1D1*: upregulation, 1.56%; homozygous deletion (Hom Del), 0.17%; unaltered, 98.27%. *NR1D2*: upregulation, 4.54%; unaltered, 95.46%. **g**, *In vivo* treatment with SR9009 results in the decrease of ATG7 protein levels (6 days, 200 mg kg^{-1} twice a day, *n* = 5 mice, one experiment). **h**, SR9009 treatment impairs *in vivo* growth of glioblastoma patient-derived xenografts (6 days, 200 mg kg^{-1}, *n* = 5 mice). **i**, Quantification of tumour size by *in vivo* luciferase assays (mean ± s.e.m., *n* = 10 mice, one-tailed Mann–Whitney test, ***P* = 0.0057). **j**, SR9009 improves survival in mice that bear glioblastoma patient-derived xenografts. SR9009 200 mg kg^{-1} twice a day, *n* = 11 (vehicle), *n* = 11 (SR9009), *n* = 11 (temozolomide (82.5 mg kg^{-1} once a day for 5 days)) mice; two-tailed log-rank analyses. For gel source data, see Supplementary Fig. 1.

doi:10.1038/nature25174

AMD1 mRNA employs ribosome stalling as a mechanism for molecular memory formation

Martina M. Yordanova[1]*, Gary Loughran[1]*, Alexander V. Zhdanov[1], Marco Mariotti[1,2], Stephen J. Kiniry[1], Patrick B. F. O'Connor[1], Dmitry E. Andreev[1,3], Ioanna Tzani[1], Paul Saffert[1], Audrey M. Michel[1], Vadim N. Gladyshev[2], Dmitry B. Papkovsky[1], John F. Atkins[1,4] & Pavel V. Baranov[1]

In addition to acting as template for protein synthesis, messenger RNA (mRNA) often contains sensory sequence elements that regulate this process[1,2]. Here we report a new mechanism that limits the number of complete protein molecules that can be synthesized from a single mRNA molecule of the human *AMD1* gene encoding adenosylmethionine decarboxylase 1 (AdoMetDC). A small proportion of ribosomes translating *AMD1* mRNA stochastically read through the stop codon of the main coding region. These readthrough ribosomes then stall close to the next in-frame stop codon, eventually forming a ribosome queue, the length of which is proportional to the number of AdoMetDC molecules that were synthesized from the same *AMD1* mRNA. Once the entire spacer region between the two stop codons is filled with queueing ribosomes, the queue impinges upon the main *AMD1* coding region halting its translation. Phylogenetic analysis suggests that this mechanism is highly conserved in vertebrates and existed in their common ancestor. We propose that this mechanism is used to count and limit the number of protein molecules that can be synthesized from a single mRNA template. It could serve to safeguard from dysregulated translation that may occur owing to errors in transcription or mRNA damage.

AdoMetDC catalyses the decarboxylation of *S*-adenosylmethionine (AdoMet or SAM), which provides an aminopropyl group to polyamines such as spermidine and spermine[3]. SAM is an important metabolite because it serves as a major donor of methyl groups in numerous reactions that involve methylation of DNA, RNA, proteins and metabolites[4–6]. Thus, in addition to its essential role in polyamine synthesis, AdoMetDC may also influence methylation reactions by affecting SAM availability. AdoMetDC is critical for embryonic stem cell self-renewal and differentiation to the neural lineage[7]. Its dysregulation is linked to tumorigenesis[8], and its overexpression in rodent fibroblasts gives rise to aggressive transformants with extremely high invasive capacity in nude mice[9]. The synthesis of AdoMetDC is tightly controlled at the translational level, allowing for quick adjustment in response to changes in polyamine concentrations. In vertebrates, control is dependent on translation of an upstream open reading frame (uORF) encoding the micropeptide MAGDIS[10]. MAGDIS stalls ribosomes at the uORF stop codon; the duration of the ribosome arrest depends on the concentration of polyamines. At high concentrations, extended pausing of ribosomes at the MAGDIS uORF inhibits translation initiation at the downstream ORF encoding AdoMetDC. At lower concentrations, ribosomes terminate at the uORF stop codon and can then efficiently reinitiate translation at the AdoMetDC ORF. This mechanism provides a negative feedback control loop for AdoMetDC autoregulation[10].

Analysis of publicly available ribosome profiling data[11] confirms translation of the MAGDIS uORF and reveals that it has the highest density of ribosome-protected fragments within the *AMD1* mRNA (Fig. 1a). However, the ribosome density profile of the *AMD1* mRNA also revealed an unexpected feature: a strong isolated peak of ribosome footprint density in its 3′ trailer (also known as 3′ untranslated region), 384 nucleotides downstream of the *AMD1* stop codon (Fig. 1a). In general, prominent isolated peaks of ribosome footprint density are indicative of an extended translational pause but could instead result from mRNA protection not related to genuine translation, for example within a nucleoprotein complex with similar sedimentation properties to that of ribosomes. The latter was recently proposed as an explanation for the peak in the *AMD1* 3′ trailer[12]. The former could potentially occur if ribosomes read through the stop codon of the annotated *AMD1* coding sequence (CDS) and then stall downstream. Occurrences of efficient stop codon readthrough have been observed in the decoding of many viruses, and in cellular genes of many organisms, including humans[13–15]. There are no in-frame stop codons downstream of the *AMD1* stop until the peak of high density that occurs at the next in-frame stop codon. Thus, stop codon readthrough is a plausible explanation for the observed ribosome density peak.

Evidence supportive of functionally significant stop codon readthrough was obtained by phylogenetic analysis of the downstream ORF, which we refer to hereafter as the *AMD1* tail, and its product as the AdoMetDC extension. We initially examined a University of California, Santa Cruz 100-species genomic alignment[16] with CodAlignView (Supplementary Data 1). The *AMD1* tail ORF appeared to be conserved in the genome from approximately 80 tetrapods. Given the low quality of some genomic sequences, we expanded our analysis to all available vertebrate genomes and applied filtering to improve the quality (see Methods). This produced an alignment of 146 species (Supplementary Data 2). The origin of the *AMD1* tail dates to at least the root of vertebrates although it has been lost in a small group of bony fish (gobies). The level of nucleotide conservation in the *AMD1* tail is similar to that of the *AMD1* main ORF, peaking towards the middle. Analysis of the non-synonymous/synonymous substitutions ratio (K_a/K_s ratio) revealed weak purifying selection acting on amino acids encoded by the *AMD1* tail, which is strongest towards the end (last ~20 amino acids: see Fig. 1b).

Examination of all available vertebrate ribosome profiling data in the GWIPS-viz browser[11] confirmed the existence of a ribosome density peak at the 3′ end of the *AMD1* tail in mouse, rat, frog and fish (Extended Data Fig. 1), strongly suggesting evolutionary conservation not only of the tail ORF, but also of the potential ribosome stalling at its 3′ end. The ribosome density peak at the end of the *AMD1* tail is also present in ribosome profiling data obtained from cells treated with drugs that preferentially block ribosomes at sites of translation initiation[17,18]. This is characteristic of ribosome stalling sites and a similar ribosome peak can be observed for the well-characterized stalling site at the end of the *XBP1* coding region[19] (Extended Data Fig. 2).

[1]School of Biochemistry and Cell Biology, University College Cork, Cork T12 YN60, Ireland. [2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02215, USA. [3]Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow 119234, Russia. [4]Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA.
*These authors contributed equally to this work.

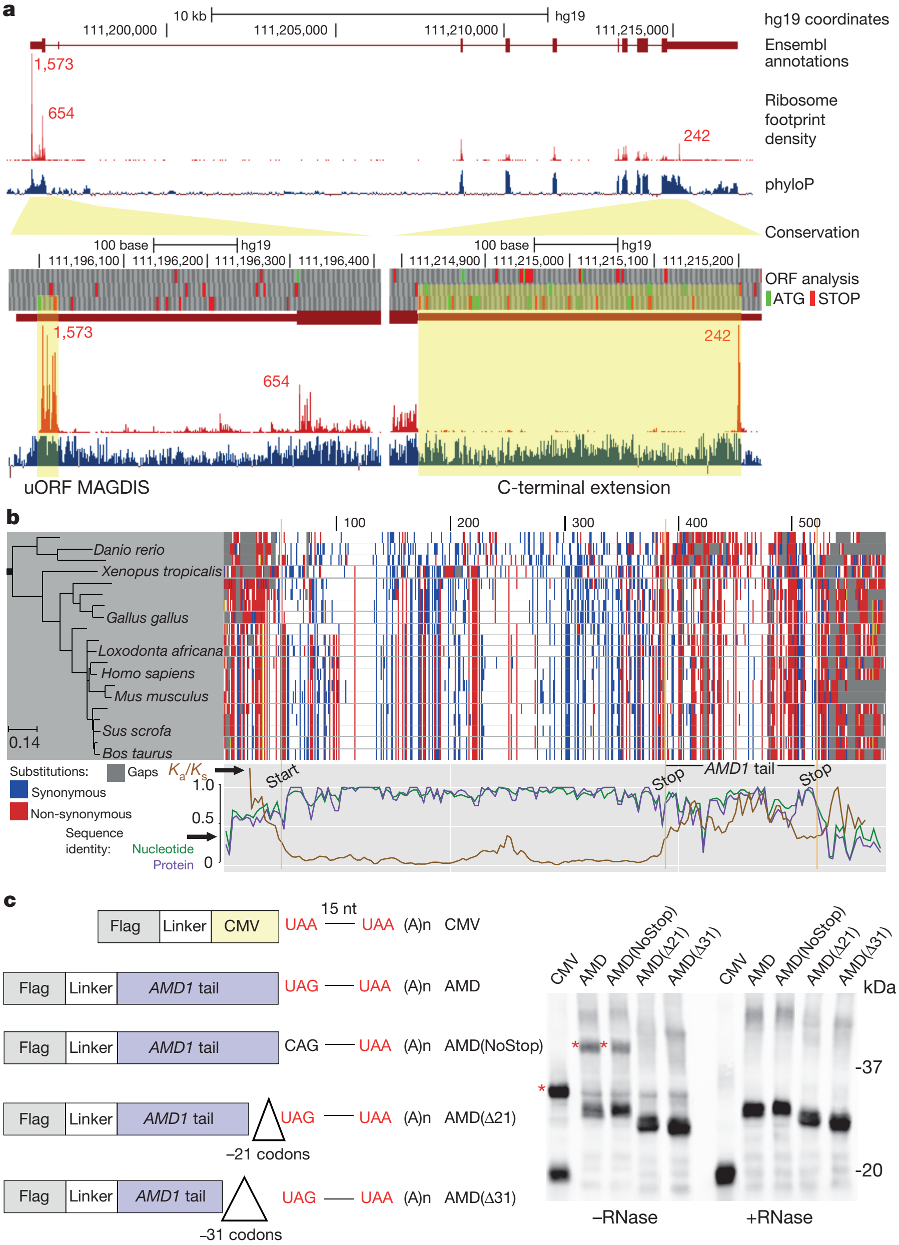


Figure 1 | Translation of phylogenetically conserved *AMD1* tail results in ribosome stalling. a, Ribosome footprint density from aggregated data-tracks in GWIPS-viz. Top plot shows all footprints in the *AMD1* locus of the human genome (hg19). Bottom plots show magnifications of the MAGDIS uORF (left) and the *AMD1* stop codon (right) areas. Footprint numbers at specific peaks are indicated in red; phyloP score indicates nucleotide conservation. **b**, Alignment of AMD1 sequences, coloured to indicate substitutions and gaps. K_a/K_s ratios and sequence identities are shown at the bottom. **c**, Immunoblotting of cell-free translated Flag-tagged *AMD1* tails (or human cytomegalovirus (CMV) uORF2) showing RNase-sensitive covalent complexes of peptidyl–tRNA; $n = 2$. AMD(NoStop), *AMD1* tail with the UAG stop codon changed to a CAG sense codon; AMD(Δ21), *AMD1* tail with the last 21 codons removed; AMD(Δ31), *AMD1* tail with the last 31 codons removed; nt, nucleotide.

To determine whether the observed *AMD1* tail peak occurs as a result of arrested ribosomes or because of protection by an RNA-binding protein as suggested earlier[12], we tested for ribosome stalling by monitoring the formation of stable peptidyl–tRNA (transfer RNA) complexes[20] during translation of full-length and truncated *AMD1* tails (Fig. 1c). Translation of *AMD1* tails resulted in the formation of RNase-sensitive complexes that depended on the 63 nucleotides immediately 5′ of the *AMD1* tail stop, although the stop codon itself was not essential (Fig. 1c). Therefore, ribosomes are stalled in a sequence-specific but stop codon-independent manner at the 3′ end of the *AMD1* tail ORF.

In an attempt to verify *AMD1* stop codon readthrough and concomitant translation of the *AMD1* tail, we transfected cells with N-terminally HA-tagged AdoMetDC (Fig. 2a). Stop codon readthrough would be expected to yield an additional C-terminally extended product, longer by 128 residues. No readthrough product was detected by western blotting using HA-tag antibodies (Fig. 2a). Surprisingly, however, we observed an almost complete loss of product in a readthrough-positive control where the *AMD1* stop codon was replaced with a sense codon (UGG construct). These observations suggest that translation of the *AMD1* tail leads to a dramatic decrease in the corresponding protein product levels. To quantify this effect, the *AMD1* tail was cloned into a dual luciferase vector downstream of, and in-frame with, sequence encoding *Renilla* luciferase (R Luc). Translation of downstream internal control firefly luciferase (F Luc) was initiated at an introduced encephalomyocarditis virus (EMCV) internal ribosome entry site (IRES) (Fig. 2b), allowing monitoring of both luciferase activities from the same mRNA. We observed an approximately 65-fold reduction in relative R Luc activity in those constructs in which the UGA stop codon

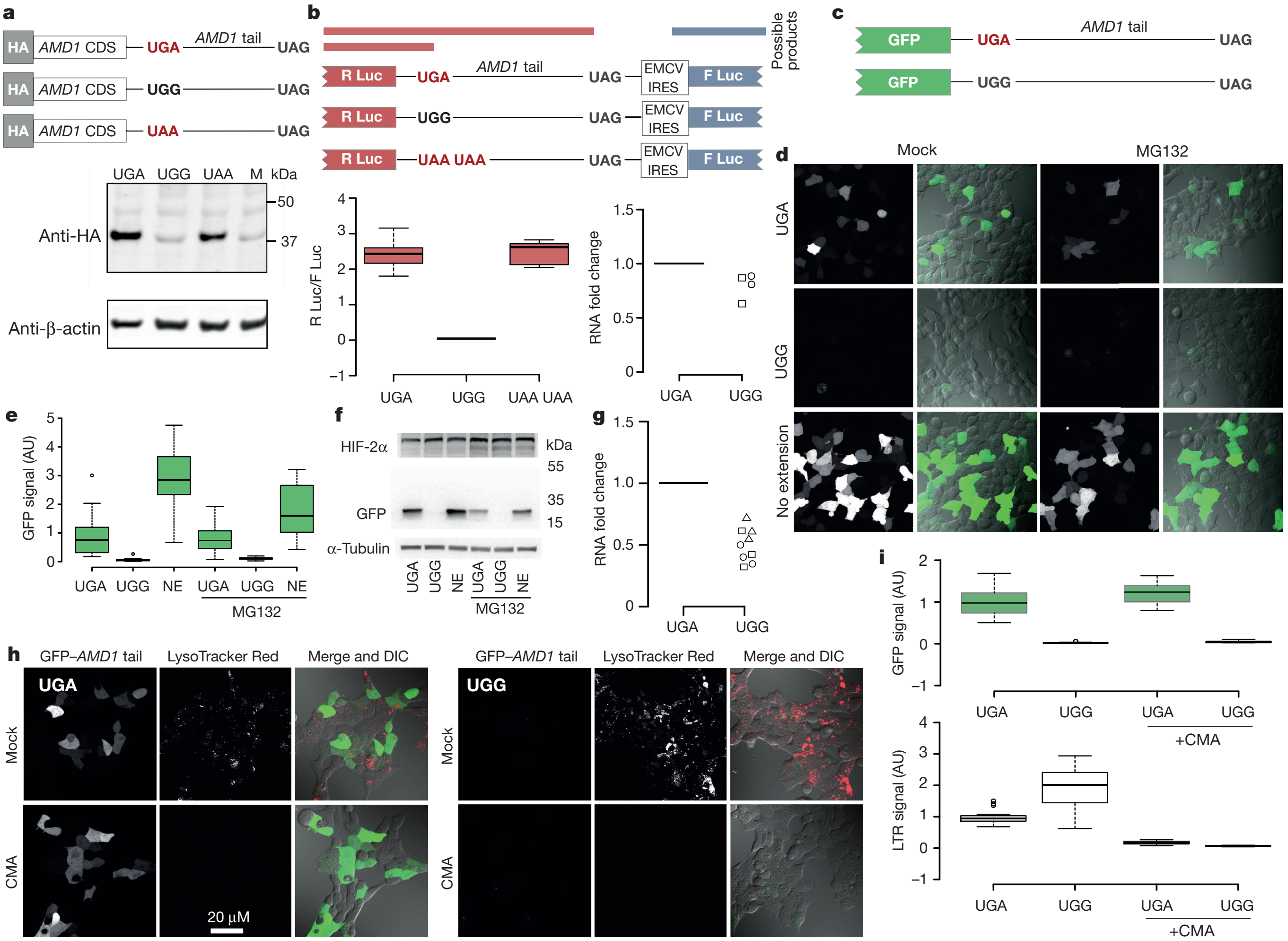


Figure 2 | *AMD1* tail translation reduces expression independent of proteasome or lysosome pathways. a, Immunoblotting of protein lysates from HEK293T cells transfected with indicated constructs (M, mock; HA, haemagglutinin); $n = 4$. **b**, Analysis of *AMD1* tail translation effect on expression of dual luciferase reporter (top); luciferase activity (bottom left; $n = 12$), mRNA stability (bottom right; $n = 4$). **c**, GFP constructs used in **d**–**i**. Imaging (**d**, **e**; $n = 16$), immunoblotting (**f**, $n = 1$) and RT–qPCR (**g**, $n = 9$) analysis of cells expressing GFP constructs with or without MG132 (10 μM); AU, arbitrary units. **h**, **i**, Imaging of cells expressing GFP constructs with or without concanamycin A (CMA; 1 μM). Acidic compartments were stained with LysoTracker Red; $n = 15$). DIC, differential interference contrast microscopy; NE, no extension.

was replaced by a UGG codon (Fig. 2b). IRES-driven F Luc activity was similar from both constructs, arguing against the possibility that *AMD1* tail translation and ribosome stalling affect mRNA stability (Extended Data Fig. 3). However, to further exclude this possibility, we measured reporter mRNA levels using quantitative PCR with reverse transcription (RT–qPCR) and did not observe changes that could explain the large reduction in the UGG construct product (Fig. 2b and Extended Data Fig. 3).

The reduction in protein levels when the *AMD1* tail is translated could be explained by an effect of the extension on AdoMetDC (and reporter) stability, localization or translation. Indeed, protein destabilization via readthrough extension was reported previously for the yeast *PDE2* mRNA[21] and was recently proposed to be a general phenomenon[22]. To explore the effect of the extension on protein stability, we designed additional constructs in which the AdoMetDC extension was fused to the C terminus of GFP (Fig. 2c). Live-cell confocal imaging and immunoblotting demonstrated a dramatic decrease in GFP levels in cells transfected with the UGG construct (Fig. 2d–f) that could not be explained by the twofold reduction in the UGG mRNA levels (Fig. 2g). The marked GFP reduction was also observed in constructs where UGA was replaced with sense codons other than UGG (Extended Data Fig. 4). The *AMD1* tail inhibitory effect was consistent across several reporter constructs, indicating that its effect is independent of the main *AMD1* coding sequence.

Treatment of cells with the proteasome inhibitor MG132 did not increase the levels of GFP–UGG (Fig. 2d–f). Similarly, GFP–UGG levels remained low upon lysosome inhibition by dissipating their acidic pH with concanamycin A (Fig. 2h, i). Cell-free translation of the luciferase reporters in HEK293T lysates again resulted in substantial reductions in R Luc levels from UGG mRNA compared with UGA mRNA, thus excluding the possibility that the observed reduction is due to a non-canonical secretion signal in the AdoMetDC extension (Extended Data Fig. 5). Furthermore, we excluded the possibility that the *AMD1* tail effect may be a result of the *AMD1* tail translation acting in *trans* (Extended Data Fig. 6).

If the AdoMetDC extension does not affect protein stability and does not serve as a secretion signal, what could be responsible for the reduction in reporter expression? To explain our observations, we propose the model illustrated in Fig. 3a. Infrequent readthrough ribosomes encounter a strong stalling site close to the *AMD1* tail stop codon and form a stable complex with the mRNA. Trailing ribosomes that also read through the *AMD1* stop codon are stymied upstream and form a queue. The length of the queue is expected to be proportional to the number of AdoMetDC molecules produced from the same mRNA. Once the entire *AMD1* tail is filled with queued ribosomes, all ribosomes translating *AMD1* would be unable to finish synthesis of AdoMetDC unless the roadblock of queueing ribosomes is cleared. Such a mechanism could be used as a safeguard against dysregulated

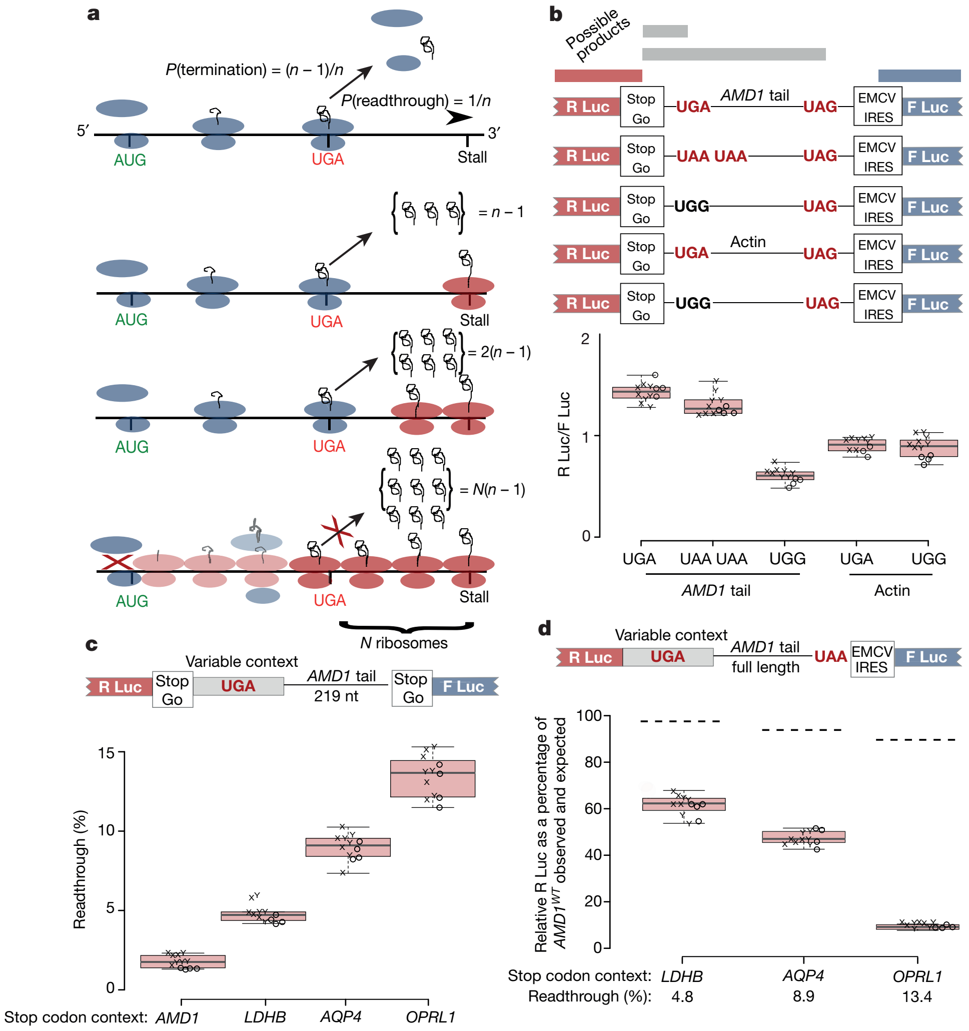


Figure 3 | *AMD1* tail modulates translation of an upstream ORF. **a**, Schematic of the proposed mechanism; stalled ribosomes are in red. The maximum number of proteins produced from a single mRNA molecule is proportional to the length of the tail (*N* ribosomes). **b**, Relative luciferase activities of indicated reporters showing the *AMD1* tail translation effect on upstream reporter; $n = 12$. **c**, Readthrough efficiencies determined by dual luciferase assay from indicated constructs; $n = 12$. **d**, The effect of stop codon readthrough context on upstream reporter activity; $n = 12$. Broken lines indicate expected relative R Luc activities if extended proteins were immediately degraded. WT, wild type.

AMD1 mRNA molecules, for example those where, owing to a synthesis error or damage, uORF-mediated repression does not work properly. Translation of such dysregulated molecules would stop after a defined number of AdoMetDC molecules had been synthesized (Fig. 3a).

To test this hypothesis, we first took advantage of the StopGo peptide motif (also known as Stop-CarryOn or 2A), which effectively results in the skipping of a peptide bond by causing release of a nascent peptide in the absence of a stop codon and then continued translation[23,24]. We fused StopGo sequences to the 3′ end of R Luc sequences before the stop codon (or sense codon control) and the *AMD1* tail (Fig. 3b). Therefore, the amino-acid sequences of R Luc reporters produced from these constructs should be identical irrespective of whether the *AMD1* tail is translated. Despite this, we consistently observed a threefold reduction in relative R Luc activity in the construct where the wild-type UGA codon was substituted with UGG (Fig. 3b). Western blotting analysis of R Luc and F Luc confirmed that the luciferase products from both constructs are identical in size (Extended Data Fig. 7). This reduction in R Luc levels cannot be explained by an effect of the AdoMetDC extension on protein product properties (owing to the presence of StopGo) or by mRNA stability, since the levels of IRES-driven F Luc observed with the UGA and UGG constructs were comparable (also confirmed by RT–qPCR; Extended Data Fig. 7c). Relative R Luc activities in UGG constructs with StopGo (Fig. 3b) were reduced less than constructs without it (Fig. 2b), probably because StopGo is a slow process and it may decelerate queue formation. We also cannot completely rule out the partial involvement of an uncharacterized protein degradation pathway.

According to our model, an increased readthrough efficiency should accelerate formation of the queue, intensifying concomitant reduction in reporter expression. To test this, we measured the activity of R Luc in constructs containing the *AMD1* tail with R Luc stop codons in contexts known to permit varying levels of readthrough[14]. To measure readthrough efficiencies of the different stop codon contexts (from *LDHB*, *AQP4*, *OPRL1* genes), it was necessary to first eliminate the *AMD1* tail effect by removing the last 50 codons (Fig. 3c), which included the sequence essential for stalling (Fig. 1c). A readthrough efficiency of approximately 1.6% was observed for the wild-type *AMD1* context (Fig. 3c), consistent with the low footprint density observed in ribosome profiles on the *AMD1* tail (Fig. 1a). The readthrough levels observed at other contexts were approximately 5%, 9% and 13.5% (Fig. 3c). As predicted from the model in Fig. 3a, reductions in relative R Luc activity were much greater than what would be expected solely because of inactivation of readthrough products when the stalling site is present (Fig. 3d).

The scheme shown could be a simplification of the real situation. Provided that the stalled ribosomes are released with a certain rate *s*, ribosomes would accumulate in the *AMD1* tail only if $s < i/n$, where *i* is the rate of initiation and $1/n$ is the probability of stop codon readthrough. In this case, the proposed mechanism would be predicted to block translation on only those mRNA molecules at which the synthesis of AdoMetDC exceeds a certain rate ($i > s/n$). Such regulation would be particularly effective in *AMD1* mRNA since the AdoMetDC half-life is less than 1 h (ref. 25) and its cellular concentration is largely determined by its synthesis. It is conceivable that similar mechanisms regulate expression of other genes where tight control is required. Indeed, we identify several genes with ribosome footprint peaks between the protein-coding ORF stop and the next in-frame stop codon (Supplementary Data 3), which, intriguingly, contains *EEF1A2* (Extended Data Fig. 8). The exact function of stalling following stop codon readthrough needs to be investigated in a case-by-case manner, as it may vary. Queueing at the end of yeast antizyme (*oaz*) ORF, for example, has been reported to reduce the efficiency of programmed ribosomal frameshifting in a polyamine-dependent manner[26]. However, because the formation of the long queues that are required for both proposed models has not been observed directly, the possibility of alternative mechanisms responsible for long-range coordination between stalled ribosomes and translation far upstream on the same mRNA cannot be excluded.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 June 2016; accepted 28 November 2017.
Published online 3 January 2018.

1. Gebauer, F. & Hentze, M. W. Molecular mechanisms of translational control. *Nat. Rev. Mol. Cell Biol.* **5,** 827–835 (2004).
2. Sonenberg, N. & Hinnebusch, A. G. New modes of translational control in development, behavior, and disease. *Mol. Cell* **28,** 721–729 (2007).
3. Pegg, A. E. *S*-Adenosylmethionine decarboxylase. *Essays Biochem.* **46,** 25–46 (2009).
4. Chiang, P. K. *et al.* S-Adenosylmethionine and methylation. *FASEB J.* **10,** 471–480 (1996).
5. Lu, S. C. & Mato, J. M. S-adenosylmethionine in liver health, injury, and cancer. *Physiol. Rev.* **92,** 1515–1542 (2012).
6. Roje, S. *S*-Adenosyl-L-methionine: beyond the universal methyl group donor. *Phytochemistry* **67,** 1686–1698 (2006).
7. Zhang, D. *et al.* AMD1 is essential for ESC self-renewal and is translationally down-regulated on differentiation to neural precursor cells. *Genes Dev.* **26,** 461–473 (2012).
8. Scuoppo, C. *et al.* A tumour suppressor network relying on the polyamine–hypusine axis. *Nature* **487,** 244–248 (2012).
9. Paasinen-Sohns, A. *et al.* Chaotic neovascularization induced by aggressive fibrosarcoma cells overexpressing S-adenosylmethionine decarboxylase. *Int. J. Biochem. Cell Biol.* **43,** 441–454 (2011).
10. Law, G. L., Raney, A., Heusner, C. & Morris, D. R. Polyamine regulation of ribosome pausing at the upstream open reading frame of *S*-adenosylmethionine decarboxylase. *J. Biol. Chem.* **276,** 38036–38043 (2001).
11. Michel, A. M. *et al.* GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res.* **42,** D859–D864 (2014).
12. Ji, Z., Song, R., Huang, H., Regev, A. & Struhl, K. Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat. Biotechnol.* **34,** 410–413 (2016).
13. Schueren, F. *et al.* Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals. *eLife* **3,** e03640 (2014).
14. Loughran, G. *et al.* Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res.* **42,** 8928–8938 (2014).
15. Stiebler, A. C. *et al.* Ribosomal readthrough at a short UGA stop codon context triggers dual localization of metabolic enzymes in fungi and animals. *PLoS Genet.* **10,** e1004685 (2014).
16. Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43,** D670–D681 (2015).
17. Gao, X. *et al.* Quantitative profiling of initiating ribosomes *in vivo*. *Nat. Methods* **12,** 147–153 (2015).
18. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).
19. Yanagitani, K. *et al.* Cotranslational targeting of XBP1 protein to the membrane promotes cytoplasmic splicing of its own mRNA. *Mol. Cell* **34,** 191–200 (2009).
20. Yanagitani, K., Kimata, Y., Kadokura, H. & Kohno, K. Translational pausing ensures membrane targeting and cytoplasmic splicing of *XBP1u* mRNA. *Science* **331,** 586–589 (2011).
21. Namy, O., Duchateau-Nguyen, G. & Rousset, J. P. Translational readthrough of the PDE2 stop codon modulates cAMP levels in *Saccharomyces cerevisiae*. *Mol. Microbiol*. **43,** 641–652 (2002).
22. Arribere, J. A. *et al.* Translation readthrough mitigation. *Nature* **534,** 719–723 (2016).
23. Ryan, M. D. & Drew, J. Foot-and-mouth disease virus 2A oligopeptide mediated cleavage of an artificial polyprotein. *EMBO J.* **13,** 928–933 (1994).
24. Doronina, V. A. *et al.* Site-specific release of nascent chains from ribosomes at a sense codon. *Mol. Cell. Biol.* **28,** 4227–4239 (2008).
25. Miller-Fleming, L., Olin-Sandoval, V., Campbell, K. & Ralser, M. Remaining mysteries of molecular biology: the role of polyamines in the cell. *J. Mol. Biol.* **427,** 3389–3406 (2015).
26. Kurian, L., Palanimurugan, R., Gödderz, D. & Dohmen, R. J. Polyamine sensing by nascent ornithine decarboxylase antizyme stimulates decoding of its mRNA. *Nature* **477,** 490–494 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to I. Jungreis and M. Kellis for allowing us to use CodAlignView and to P.R. Bhatt for technical advice on stable peptidyl–tRNA complex formation. We acknowledge financial support from Science Foundation Ireland ((12/IA/1335) to P.V.B., (13/IA/1853) to J.F.A. and (12/RC/2276) to D.B.P.); National Institute of Health (CA080946) and (GM065204) to V.N.G.; Health Research Board (PhD/2007/04) to I.T.; and Russian Science Foundation (16-14-10065) to D.E.A.

Author Contributions P.B.F.O. made the initial observation of unusual ribosome footprint density. M.M. and P.V.B. performed phylogenetic analysis. M.M.Y., G.L., A.V.Z., I.T., P.S. and D.E.A. designed and performed biochemical experiments. S.J.K. and A.M.M. analysed publicly available ribosome profiling data. P.V.B. proposed the model. M.M.Y. and P.V.B. drafted the manuscript. V.N.G., D.B.P. and J.F.A. contributed to interpretation of the experimental data and editing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to P.V.B. (p.baranov@ucc.ie).

Reviewer Information *Nature* thanks A. Geballe, P. Van Damme and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Cloning. Oligonucleotides were synthesized by IDT (Belgium). Primer sequences are listed in Supplementary Information. The sequence of the *AMD1* coding region and the part of the 3′ trailer encoding the tail was obtained as a gBlock from IDT and its sequence is provided in Supplementary Data 4. The amplicons were generated by standard single or multiple PCR reactions. Plasmids used in this study include pEGFP-C1 (Clontech), pcDNA3-HA (Invitrogen), pDluc[27,28] and pSGDLuc[29]. pDluc was modified such that the second luciferase reporter (firefly) is expressed under the control of the EMCV IRES. For StopGo constructs, the StopGo sequence[24] was inserted in place of the *Renilla* stop codon. For read-through measurements pSGDluc[29] was used. Construct sequences are provided in Supplementary Data 4. All constructs were transformed in *Escherichia coli* strain DH5-α and were verified by sequencing.

Tissue culture and cell treatment. Human embryonic kidney 293T cells (ATCC, tested mycoplasma negative) were maintained as monolayer cultures, grown in DMEM (Sigma-Aldrich) supplemented with 10% FBS, 1 mM L-glutamine and antibiotics at 37 °C in an atmosphere of 5% CO_2. For dual luciferase assay, 4×10^6 HEK293T cells were plated on 10-cm tissue culture dishes. After 24 h, the cells were detached with trypsin, suspended in fresh media and transfected in triplicate with Lipofectamine 2000 reagent (Invitrogen), using the 1-day protocol in which suspended cells are added directly to the DNA complexes in 96-well plates. For each transfection, the following was added to each well: 50 ng plasmid DNA, 0.4 μl Lipofectamine 2000 in 50 μl Opti-Mem (Gibco). Eighty thousand cells in 150 μl DMEM were added to the transfecting DNA complexes in each well. Transfected cells were incubated at 37 °C in 5% CO_2 for 21 h and assayed using the dual luciferase assay.

Transfections for western blotting of eGFP-encoding constructs for Extended Data Fig. 4 and for HA–*AMD1* and RT–qPCR analysis were performed in 6-well plates scaled-up from the method described for 96-well plate transfections above. The following was added to each well: 1 μg plasmid DNA, 7 μl Lipofectamine 2000 in 1 ml Opti-Mem. One million cells in 3 ml DMEM were added to the transfecting DNA complexes in each well. Transfected cells were incubated at 37 °C in 5% CO_2 for 36 h for western blotting and 21 h followed by RNA extraction for RT–qPCR.

For western blotting analysis of plasmids encoding eGFP (Fig. 2f), cells were seeded at 1×10^6 cells per well on 12-well plates and grown for 16 h before transfection. For confocal fluorescence microscopy, cells were seeded at 2×10^4 on MatTek glass-bottomed dishes pre-coated with collagen IV/poly-D-lysine. Transfection was performed using 0.2 μM of plasmid DNA per 1 cm^2, Lipofectamine 2000 and Opti-Mem for 3 h before cell treatment and staining. Loading of the cells with fluorescent indicator LysoTracker Red DND-99 (Invitrogen) (100 nM) was performed in Opti-MEM medium for 30 min in CO_2 incubator. For proteasome inhibition, 10 μM MG132 (Sigma-Aldrich) was added to transfected cells for 5 h before imaging or cell lysis for western blotting analysis. Concanamycin A (Sigma-Aldrich) treatment (1 μM) was performed for 5 h before imaging.

Confocal microscopy. Live-cell imaging was conducted on an Olympus FV1000 confocal laser scanning microscope with controlled CO_2, humidity and temperature. eGFP was excited at 488 nm (2.5–10% of laser power) with emission collected at 500–540 nm. LysoTracker Red was excited at 543 nm (15% of laser power); emission was collected at 560–600 nm. Acquisition of each spectral signal was done in sequential laser mode. Fluorescence and differential interference contrast images were collected with a 60× oil immersion objective lens in 12 planes using 0.5-μm steps. The resulting single images were analysed using FV1000 Viewer (Olympus), and Adobe Photoshop and Illustrator software. Fifteen random cells from three independent experiments were used for the box plots represented in Fig. 2.

Protein isolation and western blot analysis. For western blotting analysis of plasmids encoding eGFP for Fig. 2f, whole-cell lysates were prepared in a standard RIPA buffer (Thermo Fisher) containing protease and phosphatase inhibitors. After lysate clarification, protein concentration was measured using a BCA Protein Assay kit (Thermo Fisher) and equalized. Proteins were separated by 4–20% polyacrylamide gel electrophoresis on pre-made acrylamide gels (GenScript), transferred onto a 0.2 μm Immobilon-P PVDF membrane (Sigma-Aldrich) using wet mini-transfer system Hoefer TE 22 (Hoefer) and probed with antibodies against HIF-2α (R&D systems), α-tubulin (Sigma-Aldrich) and GFP (Novex) in 5% fat-free milk in TBST (0.8% Tween 20) overnight at 4 °C. Immunoblots were analysed with HRP-conjugated secondary antibodies (Sigma-Aldrich, 2 h at room temperature) and Amersham ECL Prime reagents using LAS-3000 Imager (Fujifilm) and Image Reader LAS-3000 2.2 software.

For eGFP-encoding plasmids for Extended Data Fig. 4, HA- and luciferase-encoding plasmids, cells were washed with 1× PBS and lysed in 1× PLB (Passive Lysis Buffer, Promega). Proteins were separated by polyacrylamide gel electrophoresis on pre-made Bolt 4–12% Bis-Tris Plus gels (Thermo Fisher), transferred onto nitrocellulose membranes (Protran) and incubated with primary antibodies in 5% fat-free milk in PBST (0.1% Tween 20) overnight at 4 °C. Primary antibodies were against HA (clone 16B12: Covance), *Renilla* (MBL), firefly (Promega) and GFP (Novex). Incubation with fluorescently labelled secondary antibodies was for 0.5 h at room temperature.

Dual luciferase assay. Firefly and *Renilla* luciferase assay buffers were prepared as described in ref. 30. Relative light units were measured on a Veritas Microplate Luminometer fitted with two injectors (Turner Biosystems). Cells transfected in 96-well plates were washed once with 1× PBS and then lysed in 18 μl of 1× PLB, and light emission was measured after injection of 50 μl of each luciferase substrate buffer.

RNA extraction and RT–qPCR. For cells transfected with dual luciferase constructs, RNA was extracted using TRIzol reagent (Ambion) according to the manufacturer's protocol followed by precipitation with isopropanol. RNA (200 ng) was DNase treated (RQ1-DNase, Promega) and reverse transcribed with Oligo dT-Primer (Fig. 2b and Extended Data Fig. 7c) or random primer (Extended Data Fig. 3b) and Superscript III (Thermo Fischer). RT–qPCR was performed in 10 μl reactions using QuantiFast SYBR Green (Qiagen). RNA levels of test constructs were normalized to β-actin (Fig. 2b and Extended Data Fig. 7c) or vimentin (Extended Data Fig. 3b) mRNA levels.

For eGFP-encoding constructs, cells were washed with 1× PBS and lysed with 1× PLB. RNA was extracted using the phenol–chloroform method and precipitated with isopropanol. RNA (1 μg) was DNase treated (RQ1-DNase). Two hundred nanograms of DNase-treated RNA were reverse transcribed with random oligonucleotides (IDT) and Superscript III (Thermo Fischer) in 20 μl reactions according to the manufacturer's recommendations. RT–qPCR was performed in 10 μl reactions using QuantiFast SYBR Green (Qiagen). RNA levels of test constructs were normalized to endogenous vimentin levels. The fold difference was calculated by the comparative C_t ($\Delta\Delta^{C_t}$) method[31]. All data points were plotted. All primer sequences are listed in the Supplementary Information.

***In vitro* transcription.** mRNA for *in vitro* translation was produced with a T7 RiboMAX Express Large Scale RNA Production System (Promega) following the manufacturer's instructions with a PCR product serving as a template (primers used to generate the PCR templates are listed in the Supplementary Information). For translation in the HEK293T cell-free system, mRNA was capped using the Vaccinia Capping System (NEB). For translation in RRL, mRNA was not capped.

***In vitro* translation in HEK293T cell-free system.** A HEK293T cell-free translation system was prepared as described in ref. 32. In brief, HEK293T cells at approximately 75% of confluence were quickly harvested on ice and resuspended in lysolecithin lysis buffer (20 mM HEPES–KOH pH 7.4, 100 mM KOAc, 2.2 mM $Mg(OAc)_2$, 2 mM DTT, 0.1 mg ml^{-1} lysolecithin). The cells were then spun down and resuspended in hypotonic extraction buffer (20 mM HEPES pH 7.5, 10 mM KOAc, 1 mM $Mg(OAc)_2$, 4 mM DTT, Complete Protease Inhibitor Cocktail (EDTA-free; Roche)). They were then disrupted in a pre-cooled 2 ml dounce homogenizer. The lysates were collected after centrifugation for 10 min at 10,000*g*. Ten-microlitre *in vitro* translation reactions were assembled in the presence of 50% v/v HEK293T cell-free lysate, 1× translation buffer (20 mM HEPES–KOH pH 7.5, 1 mM DTT, 0.5 mM spermidine–HCl, 0.6 mM Mg $(OAc)_2$, 8 mM creatine phosphate, 1 mM ATP, 0.2 mM GTP, 120 mM KOAc and 25 μM of each amino acid) and 200 ng capped mRNA, as described in ref. 33.

***In vitro* translation in RRL.** RNA was translated in 10 μl *in vitro* reactions using Flexi Rabbit Reticulocyte Lysate System (Promega) according to the manufacturer's recommendations.

Monitoring of peptidyl–tRNA complexes. Translation reactions in RRL were performed at 30 °C for 1 h and then placed on ice. Half of the reaction (5 μl) was then subjected to RNaseA treatment for 20 min on ice. RNase treated and untreated samples (2.5 μl and 2.0 μl, respectively) were combined with 2× sample buffer supplemented with RNase inhibitor and loaded onto NuPage Bis-Tris neutral gels (Thermo Fisher). The neutral pH prevents hydrolysis of the peptidyl–tRNA bond. The products were detected with an anti-Flag antibody (F1804, Sigma). The cytomegalovirus gp48 uORF2 was used as a positive control for co-translational ribosome stalling[34,35].

Gene finding and evolutionary analysis. All publicly available genome sequences of vertebrates were downloaded from NCBI. We then used Selenoprofiles[36] to identify *AMD1* orthologues using a manually curated protein profile alignment spanning the main ORF. Gene structures were completed by extending homologous coding regions to the upstream methionine and first in-frame downstream stop. Results were filtered to exclude retrotransposed pseudogenes, abundant in mammals and recognizable for their lack of introns. The predictions were further filtered through manual inspection to obtain a bona fide set of 146 complete gene sequences across vertebrates with clear orthology (Supplementary Data 5). These gene structures were then extended by 120 nucleotides upstream and 510 nucleotides downstream, to include a similar length of non-coding sequence at each side (considering the *AMD1* tail downstream of the *AMD1* main ORF stop). The phylogenetic tree of the investigated species was extracted from NCBI taxonomy[37] and standardized arbitrarily to a dichotomic tree with ETE 3 (ref. 38). Evolutionary analysis was then performed on the resulting alignment and tree using pycodeml (available at https://github.com/marco-mariotti/pycodeml). The rate of non-synonymous versus synonymous substitutions (K_a/K_s) was computed with codeml[39] using a fixed rate. This metric was computed in sliding windows throughout the alignment, each 30 codons wide and with a three-codon step. Sequence identity at nucleotide and protein levels was also computed on the same alignment in sliding windows, each three codons wide and with a three-codon step.

Codon alignment for Supplementary Data 1 was produced using CodAlignView (I. Jungreis, M. Lin and M. Kellis, manuscript in preparation).

Identification of transcripts with ribosome footprint density profiles similar to *AMD1* mRNA. For each protein coding transcript in GENCODE version 22 (ref. 40), the genomic coordinates of the region between the annotated CDS stop and the next in-frame stop were extracted. Any nucleotide position within this region that overlapped with another annotated coding region in the GENCODE annotations was discarded. The number of footprints at each remaining position was extracted from the global aggregate track of ribosome profiling data (hg38 assembly) in GWIPS-viz[11]. The positions with the highest number of footprints were recorded for each transcript. Any transcript where the peak was smaller than 500 footprints was discarded. The list of all candidates is provided in Supplementary Data 3.

The list of all datasets used in global aggregates during GWIPS-viz screenshot generation is provided in Supplementary Data 6.

Statistics and reproducibility. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Unless $n = 1$, n in all figure legends indicates the number of biological replicates from two to four independent experiments. By biological replicates, we mean measurements that were performed on distinct biological samples. By independent experiments, we mean experiments that were performed either by different investigators or took place at a different time or location.

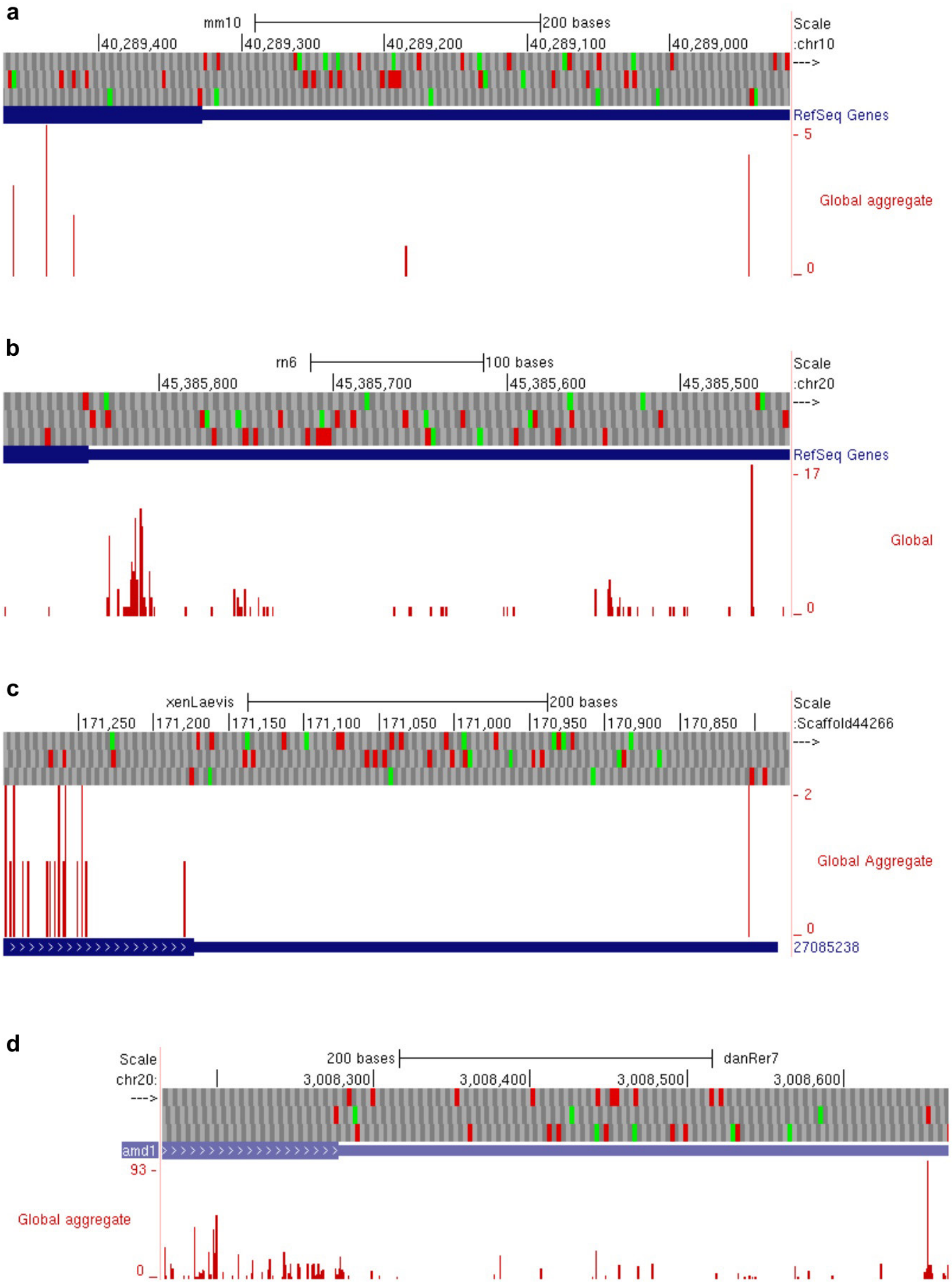
Box plots: the central line indicates the median, the box limits indicate the interquartile area, whiskers indicate 1.5 × interquartile range, and outliers (if they occur) are indicated with dots. In some of the figures, individual data points are placed on top of box plots, and data points that belong to the same biological replicate are shown with the same symbols that differ for independent experiments.

Code availability. A custom python code for finding transcripts with peaks of footprint density in ORFs extended downstream of stop codons is provided in the Supplementary Information.

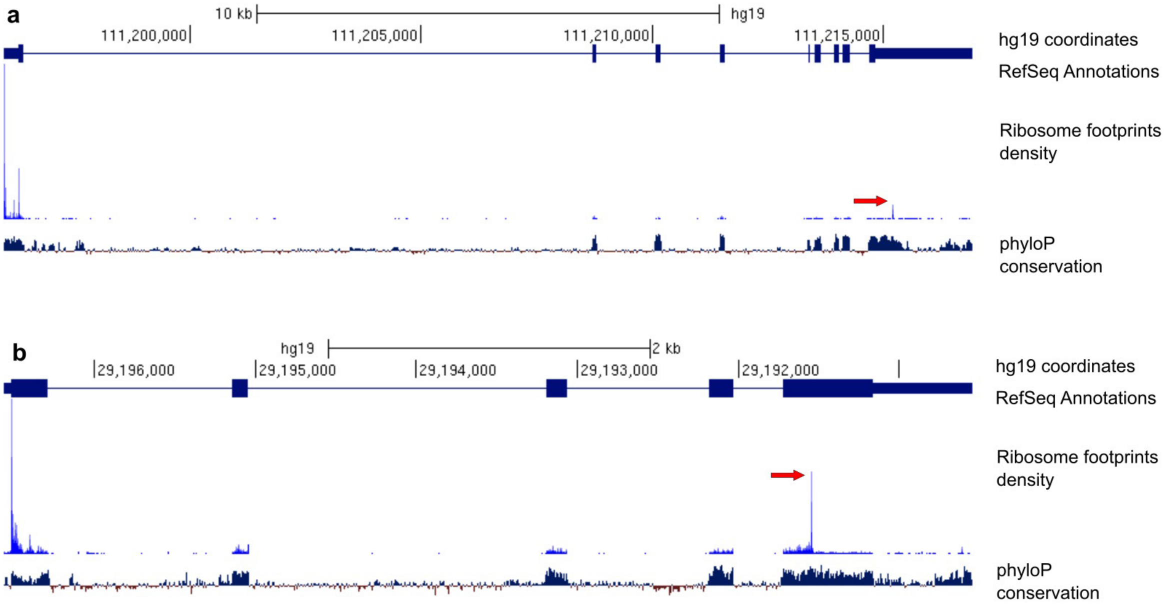
Data availability. All data generated during this study are included in this paper and its Supplementary Information. Source Data for Figs 2, 3 and Extended Data Figs 3, 6, 7 are provided in the online version of the paper; for gel source data, see Supplementary Fig. 1.

In addition, publicly available data were analysed in this study. Alignments of ribosome profiling data were obtained through GWIPS-viz (https://gwips.ucc.ie). The Gene Expression Omnibus and Sequence Read Archive accession numbers for the datasets used in GWIPS-viz for global tracks at the time of screenshots generation are available in Supplementary Data 6. Genomic sequences were obtained through GenBank, and all relevant sequences are provided in Supplementary Data 5 along with GenBank accession numbers.

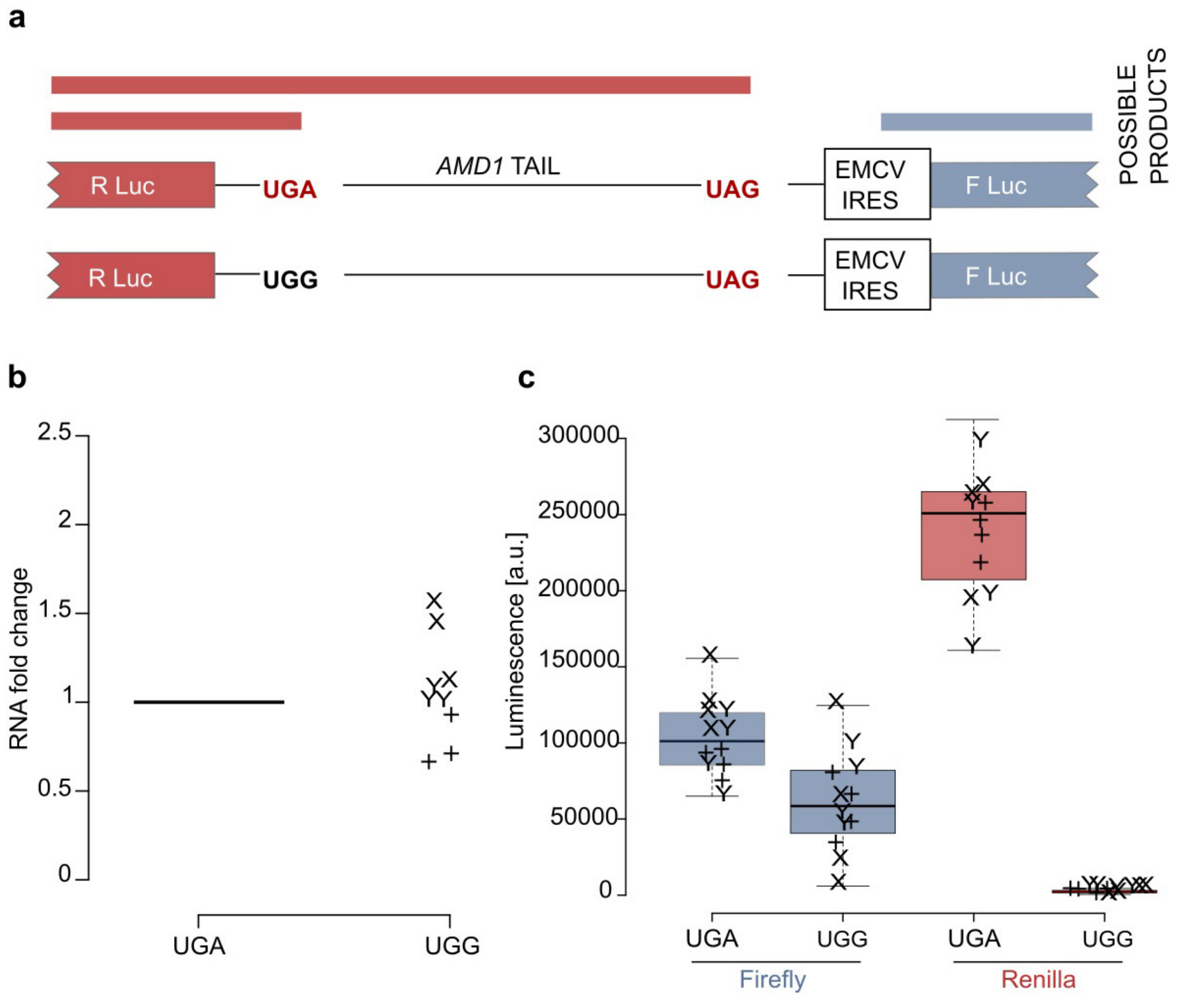
27. Fixsen, S. M. & Howard, M. T. Processive selenocysteine incorporation during synthesis of eukaryotic selenoproteins. *J. Mol. Biol.* **399,** 385–396 (2010).
28. Grentzmann, G., Ingram, J. A., Kelly, P. J., Gesteland, R. F. & Atkins, J. F. A dual-luciferase reporter system for studying recoding signals. *RNA* **4,** 479–486 (1998).
29. Loughran, G., Howard, M. T., Firth, A. E. & Atkins, J. F. Avoidance of reporter assay distortions from fused dual reporters. *RNA* **23,** 1285–1289 (2017).
30. Dyer, B. W., Ferrer, F. A., Klinedinst, D. K. & Rodriguez, R. A noncommercial dual luciferase enzyme assay system for reporter gene analysis. *Anal. Biochem.* **282,** 158–161 (2000).
31. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods* **25,** 402–408 *(*2001).
32. Terenin, I. M., Andreev, D. E., Dmitriev, S. E. & Shatsky, I. N. A novel mechanism of eukaryotic translation initiation that is neither m7G-cap-, nor IRES-dependent. *Nucleic Acids Res.* **41,** 1807–1816 (2013).
33. Andreev, D. E. *et al.* Differential contribution of the m7G-cap to the 5′ end-dependent translation initiation of mammalian mRNAs. *Nucleic Acids Res.* **37,** 6135–6147 (2009).
34. Degnin, C. R., Schleiss, M. R., Cao, J. & Geballe, A. P. Translational inhibition mediated by a short upstream open reading frame in the human cytomegalovirus gpUL4 (gp48) transcript. *J. Virol.* **67,** 5514–5521 (1993).
35. Bhushan, S. *et al.* Structural basis for translational stalling by human cytomegalovirus and fungal arginine attenuator peptide. *Mol. Cell* **40,** 138–146 (2010).
36. Mariotti, M. & Guigó, R. Selenoprofiles: profile-based scanning of eukaryotic genome sequences for selenoprotein genes. *Bioinformatics* **26,** 2656–2663 (2010).
37. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37,** D5–D15 (2009).
38. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33,** 1635–1638 (2016).
39. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
40. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).



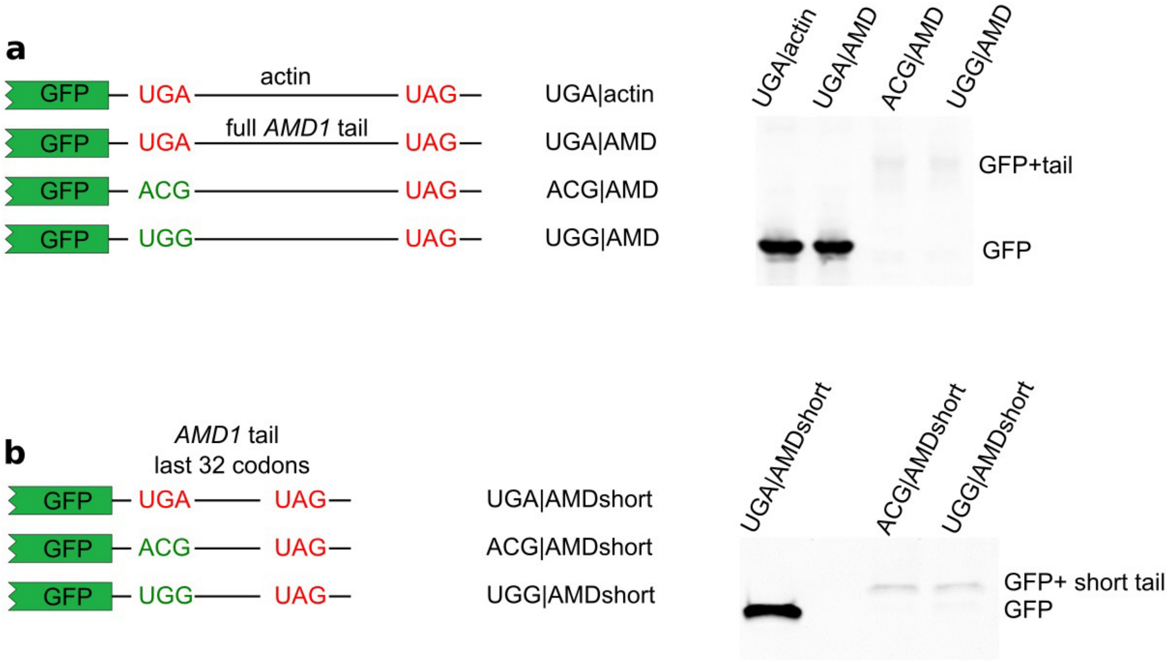
Extended Data Figure 1 | Cross-species examination of *AMD1* tail using publicly available ribosome profiling data in the GWIPS-viz browser. Available ribosome footprints aligned to the genomes of (**a**) mouse, (**b**) rat, (**c**) frog and (**d**) zebrafish are shown along with gene annotation tracks and ORF plots in which ATG codons are shown in green and stop codons are shown in red. Note that, even under very low coverage, peaks of density are consistently present at the stop codon of *AMD1* tail. The low number of footprints in mouse is due to ambiguous mapping caused by the presence of a retrotransposed single-exon AMD1 copy (AMD2). Only uniquely aligned footprints are currently displayed in GWIPS-viz.



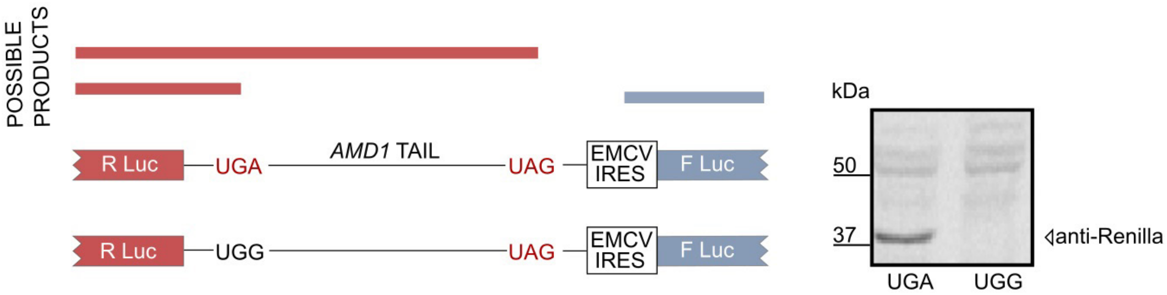
Extended Data Figure 2 | Human ribosome profiling data obtained with approaches that enrich ribosomes at translation initiation sites. **a**, *AMD1* locus; **b**, *XBP1* locus. Three tracks are shown as indicated in the figure. Ribosome footprint density corresponds to aggregated data obtained with drug treatments that preferentially arrest initiating ribosomes. Under these treatments, actively elongating ribosomes run off. However, stalled ribosomes remain bound to mRNA and produce footprints along with initiating ribosomes blocked by these inhibitors. The peaks corresponding to ribosome stalling at the end of the *AMD1* tail and at the end of the *XBP1* coding region (in unprocessed mRNA) are indicated with arrows.



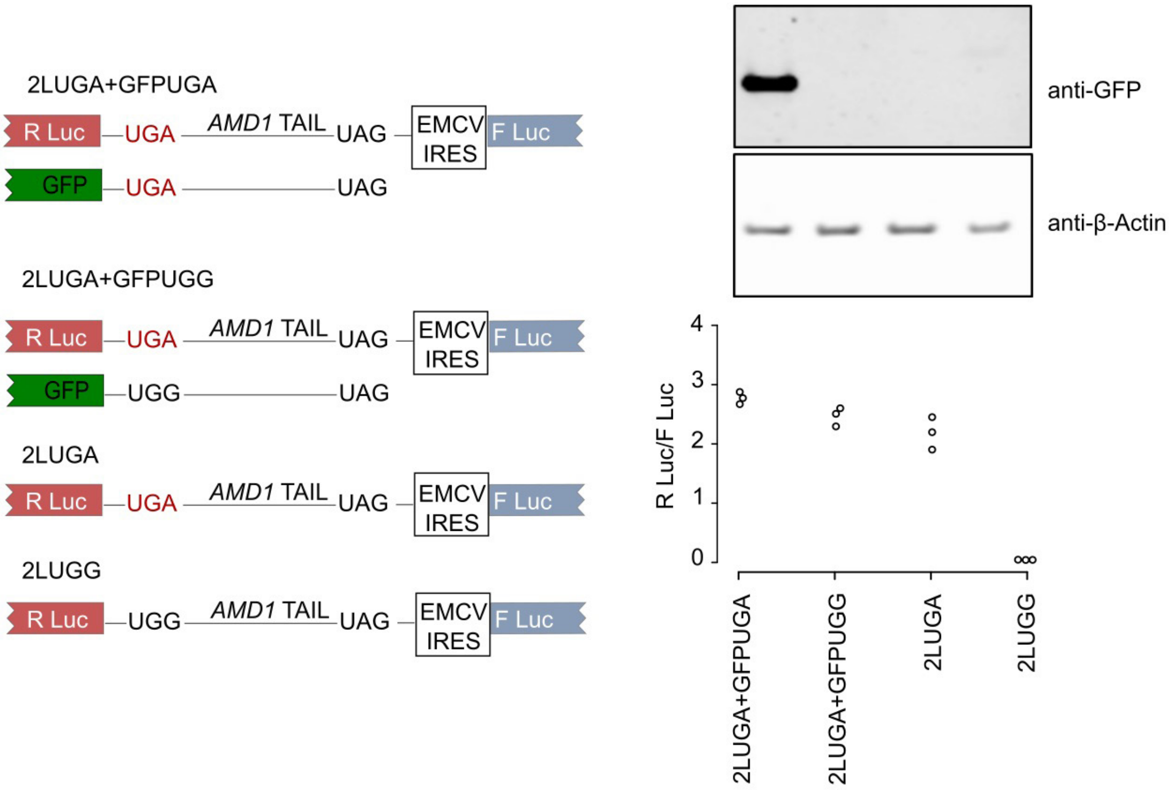
Extended Data Figure 3 | Assessment of dual luciferase mRNA stability. **a**, Scheme of constructs. **b**, RT–qPCR analysis with primers targeting R Luc sequence; *n* = 9. **c**, Absolute values of R Luc and F Luc; *n* = 12. Biological replicates that belong to the same independent experiment are indicated with the same symbols in **b** and **c**.



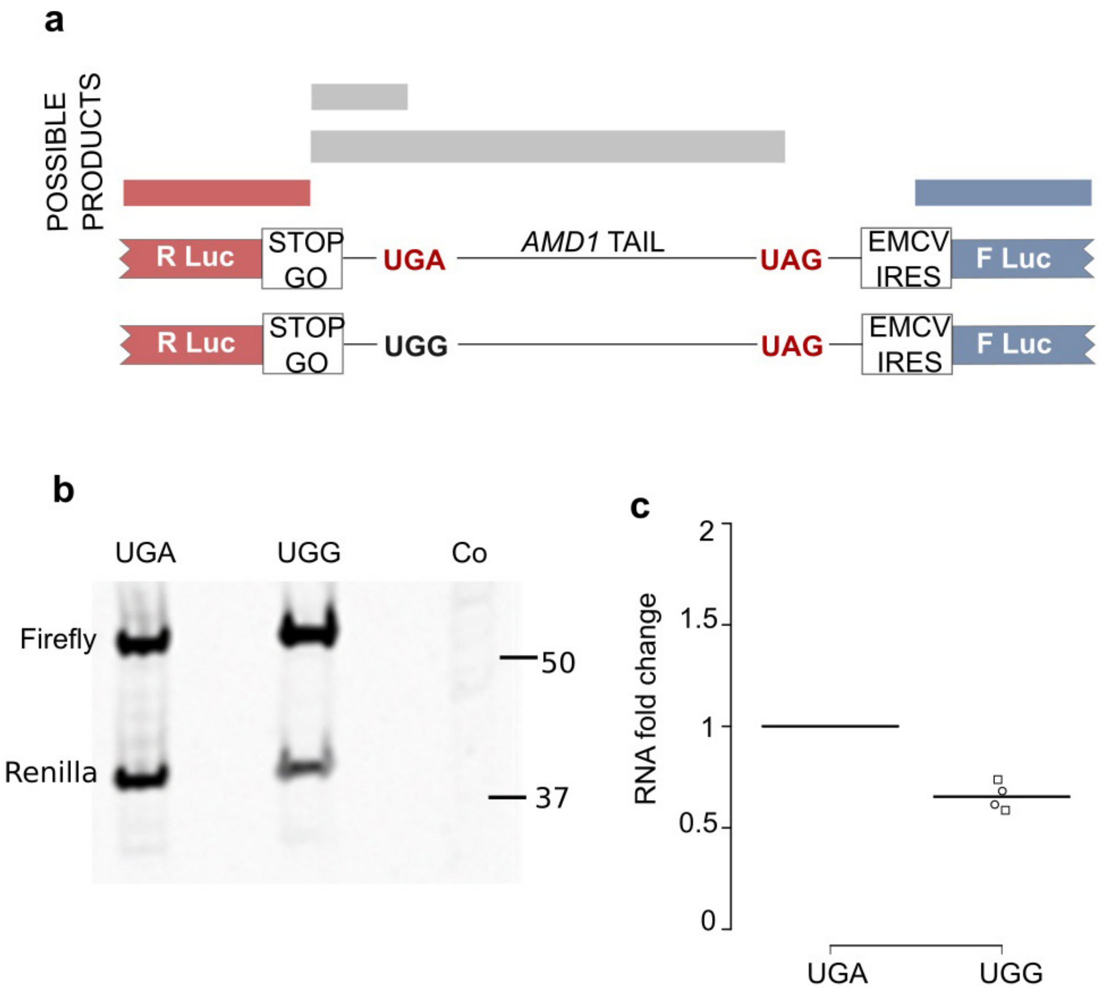
Extended Data Figure 4 | Expression of GFP constructs. Western blotting analysis of GFP fusions with a fragment of the actin 3′ trailer of the same length as full-length *AMD1* tail; $n = 2$ (**a**) and truncated from the 5′ end tail; $n = 2$ (**b**), separated by either stop or sense codons as indicated.



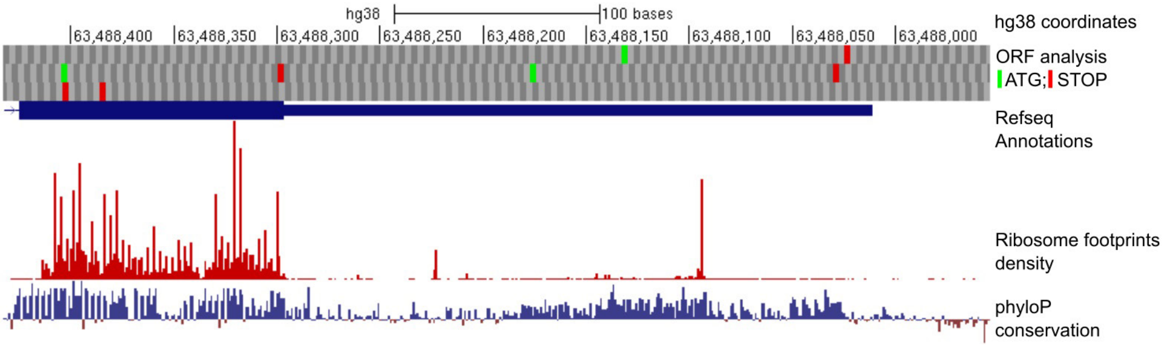
Extended Data Figure 5 | *In vitro* translation of *AMD1* tail fusion reporters. Western blotting analysis of luciferase-expressing mRNAs in the HEK293T cell-free translation system; $n = 1$.



Extended Data Figure 6 | Potential *trans* effect of *AMD1* tail translation. HEK293T cells were transfected in triplicate wells of half-area 96-well plates with the indicated expression constructs (left) for 24 h. Cells were lysed in 15 μl PLB and incubated with shaking for 15 min at room temperature. Five microlitres of each were removed for immunoblotting with both anti-GFP and anti-β-actin (upper right; $n = 3$); the remaining lysate was assayed for both R Luc and F Luc activities (lower right), $n = 3$.



Extended Data Figure 7 | Expression of StopGo constructs. a, Scheme of constructs. **b**, Western blots with antibodies against *Renilla* and firefly; $n = 1$. **c**, RT–qPCR analysis; $n = 4$. Line represents mean.



Extended Data Figure 8 | EEF1A2 readthrough extension. GWIPS-viz screenshot of ribosome footprint density at the last 3′ exon of *EEF1A2* (hg38).

doi:10.1038/nature25190

Architecture of a channel-forming O-antigen polysaccharide ABC transporter

Yunchen Bi[1], Evan Mann[2], Chris Whitfield[2] & Jochen Zimmer[1]

O-antigens are cell surface polysaccharides of many Gram-negative pathogens that aid in escaping innate immune responses[1]. A widespread O-antigen biosynthesis mechanism involves the synthesis of the lipid-anchored polymer on the cytosolic face of the inner membrane, followed by transport to the periplasmic side where it is ligated to the lipid A core to complete a lipopolysaccharide molecule[2]. In this pathway, transport to the periplasm is mediated by an ATP-binding cassette (ABC) transporter, called Wzm–Wzt. Here we present the crystal structure of the Wzm–Wzt homologue from *Aquifex aeolicus* in an open conformation. The transporter forms a transmembrane channel that is sufficiently wide to accommodate a linear polysaccharide. Its nucleotide-binding domain and a periplasmic extension form 'gate helices' at the cytosolic and periplasmic membrane interfaces that probably serve as substrate entry and exit points. Site-directed mutagenesis of the gates impairs *in vivo* O-antigen secretion in the *Escherichia coli* prototype. Combined with a closed structure of the isolated nucleotide-binding domains, our structural and functional analyses suggest a processive O-antigen translocation mechanism, which stands in contrast to the classical alternating access mechanism of ABC transporters.

Microorganisms commonly use cell surface polysaccharides to establish extended barriers that protect against the defence machineries of their hosts[1]. O-antigens help bacteria to evade innate immune responses including phagocytosis and complement-mediated lysis[3–5]. The polymers are hypervariable polysaccharides up to approximately 100 sugar units long and most reach the periplasm by one of two convergent pathways[2]. In the widespread ABC transporter-dependent pathway, the O-antigen is fully synthesized as an undecaprenyl diphosphate (UND-PP)-linked intermediate, before being transported to the periplasmic leaflet of the inner membrane by Wzm–Wzt and ligated to the lipid A core (Extended Data Fig. 1). PglK, an oligosaccharide ABC transporter from a bacterial protein *N*-glycosylation system, provided the first example of an exporter translocating UND-PP-linked substrates[6].

Some systems signal completion of O-antigen biosynthesis by modifying the growing (non-reducing) end of the polysaccharide chain with, for example, phosphate, methyl, or sugar moieties[7]. The corresponding ABC transporter recognizes the modified terminus via a carbohydrate-binding domain (CBD) fused to the C terminus of its nucleotide-binding domain (NBD) to accomplish transport[8–10] (Extended Data Fig. 1). In other systems, export of uncapped glycans, such as O-antigens and teichoic acids, occurs without the involvement of CBDs[7].

ABC transporters usually cycle between inward- and outward-facing conformations to facilitate substrate transport. However, this 'alternating access' model may not apply to transporters translocating high molecular mass polymers, such as polypeptides, O-antigens, and capsular polysaccharides.

To elucidate the O-antigen translocation mechanism, we determined the crystal structure of *A. aeolicus* (*Aa*)Wzm–Wzt, which is homologous to the prototypical *E. coli* Wzm–Wzt and *Staphylococcus aureus* wall teichoic acid transporters[11] (Extended Data Fig. 2). In a nucleotide-free state, Wzm–Wzt forms a continuous channel across the membrane. The similarity of the transporter to the *E. coli* O9a transporter allowed testing of functional predictions *in vivo* with an established prototype. Combined with structures of the transporter's isolated NBDs in a closed conformation, our structural and functional analyses suggest substrate entry and exit pathways and a model for O-antigen membrane translocation.

We initially expressed and purified the full-length *Aa*Wzm–Wzt transporter, with Wzm and Wzt forming the transmembrane domain and NBD, respectively. For crystallization, the C-terminal CBD of Wzt was removed, generating a construct including residues 1–235 (WztN). Similar constructs of *E. coli* O9a and *Klebsiella pneumoniae* O12 Wzt

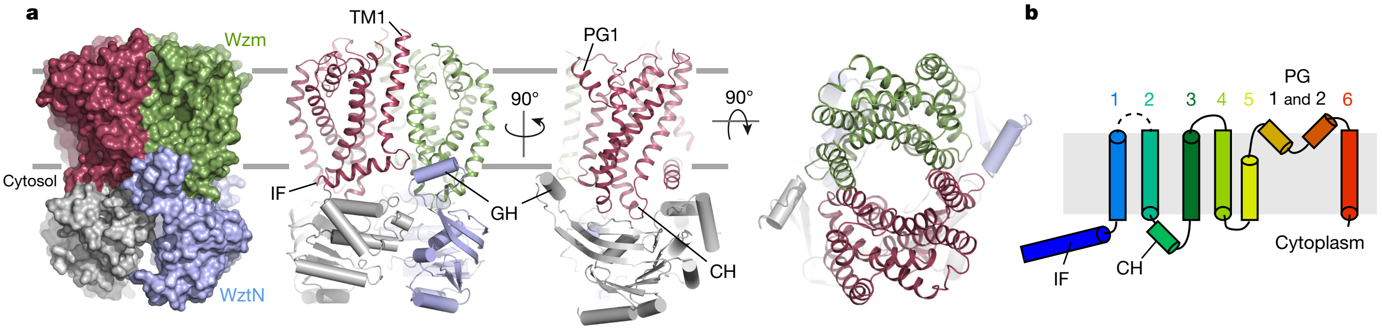


Figure 1 | Architecture of the Wzm–Wzt O-antigen transporter. a, The Wzm protomers are shown in green and red and the nucleotide-binding WztN domains are shown in blue and grey, respectively. WztN forms a short gate helix (GH) near the Wzm protomer interface and Wzm contains an N-terminal interface helix (IF). **b**, Transmembrane topology of Wzm. Wzm forms six transmembrane helices and the cytosolic TM2/3 loop forms the coupling helix (CH). The periplasmic TM5/6 loop of Wzm generates two periplasmic gate helices (PG1 and PG2). Horizontal lines indicate likely membrane boundaries.

[1]Molecular Physiology and Biological Physics, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA. [2]Department of Molecular and Cellular Biology, University of Guelph, Guelph, Ontario N1G 2W1, Canada.

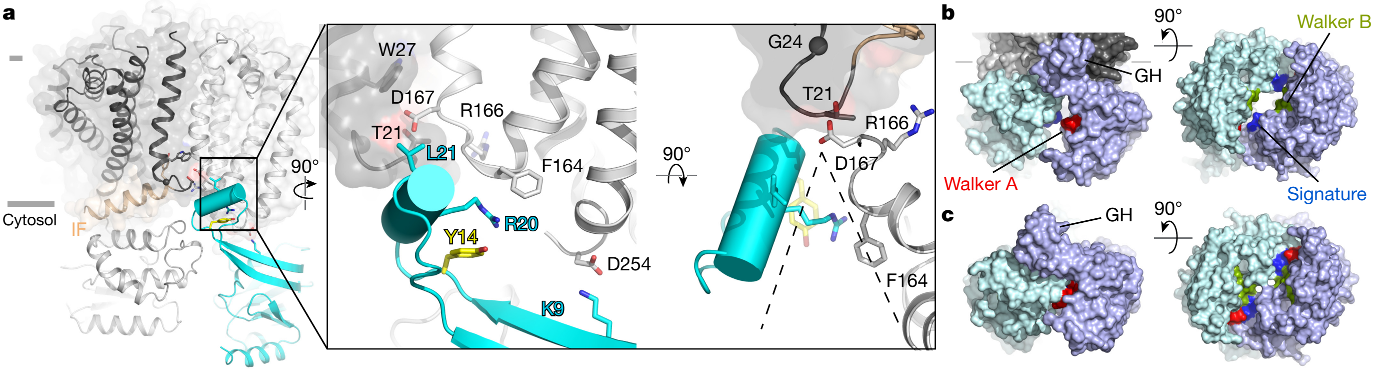


Figure 2 | Wzt forms a unique interface with Wzm. a, Position of the gate helix at the Wzm protomer interface. The GH packs against the TM4/5 loop of Wzm and forms a wedge-shaped opening towards the cytosolic water–lipid interface. The transporter is shown as a cartoon and one Wzm protomer is shown as a semi-transparent surface. **b**, Open conformation of WztN. Surface representation of the transporter's NBDs coloured blue and cyan, respectively. Conserved regions are labelled. **c**, Surface representation of the isolated WztN structure in a closed conformation coloured as in **b**.

proteins are fully functional *in vivo* if the CBD is expressed *in trans*[8,10]. The 3.85 Å-resolution Wzm–WztN structure includes residues 2–255 of Wzm and 2–235 of WztN (Extended Data Table 1 and Extended Data Fig. 3).

In a nucleotide-free state, Wzm–WztN adopts a compact structure containing a Wzm dimer interacting with a WztN dimer in an open conformation (Fig. 1a). The transmembrane domains closely interact over the entire length of their transmembrane regions and surround a central transmembrane channel, formed by transmembrane helices 1, 2, and 5, which is open to the intra- and extracellular milieu (see below).

Wzm does not contain any crossover helices that would interact with WztN of the neighbouring half-transporter, unlike previously described bacterial exporters[12]. At its N terminus, Wzm forms an amphipathic interface helix that runs parallel to the Wzm–WztN interface, followed by six transmembrane helices. The loop connecting transmembrane helices 2 and 3 (TM2 and TM3) couples Wzm with WztN (coupling helix) and the periplasmic connection between TM5 and TM6 forms two re-entrant helices, PG1 and PG2. Overall, the Wzm architecture resembles the type-II ABC exporter topology, which has so far been observed only in human lipid exporters[13,14] (Fig. 1b and Extended Data Fig. 4a). Notably, the Wzm–WztN architecture, and thus probably its translocation mechanism, differs markedly from PglK, which translocates UND-PP-linked oligosaccharides[6] (Extended Data Fig. 4b).

The NBDs of the transporter are separated by about 8 Å between the Walker A and signature motifs, sufficient for nucleotide diffusion[12] (Figs 1a and 2b). A defining feature of WztN is an extension of the β-strand 1/2 loop (residues 13–32), which forms a short 'gate helix' (residues 18–26) that rests near the Wzm–Wzm protomer interface at the putative water–lipid boundary (Fig. 2a). The gate helix packs against a loop connecting TM4 and TM5 of Wzm of the same half-transporter. This loop contains a conserved F-X-R/K-D motif, of which Phe164 interacts with Arg20 of the gate helix and Asp167 sits directly at the Wzm–Wzm interface (Fig. 2a). Additional interactions occur with backbone residues of the loop connecting the N-terminal interface helix and TM1 in the opposing Wzm subunit.

Because the C terminus of the gate helix is rotated away from the transmembrane region, the helix creates a wedge-shaped path towards the Wzm dimer interface, probably forming a substrate-binding pocket (Fig. 2a). At its centre, the gate helix contains a conserved positively charged residue (Arg20), which could be implicated in binding the UND-PP (Fig. 2a and Extended Data Fig. 2a). Strikingly, a preceding Tyr residue (Tyr14) packs against and stabilizes the gate helix on its membrane-distal side (Fig. 2a). Primary sequence alignments of homologous transporters reveal that the gate helix and the Tyr residue are characteristic features of all known O-antigen and wall teichoic acid ABC transporters (Extended Data Fig. 2a), which accept substrates synthesized as UND-PP-linked intermediates[11,15].

We also determined the structure of the isolated NBD of WztN in two different crystal forms at 2.05 and 3.5 Å resolution (Extended Data Table 1). Despite the absence of a stabilizing nucleotide, both structures represent a WztN dimer in a closed conformation, with only about 4.0 Å between the hydroxyl group of Ser61 (Walker A) and the backbone amide nitrogen of Ser143 (signature) (Fig. 2c and Extended Data Fig. 5). This closed conformation is in agreement with the adenosine 5′-(β,γ-imido)triphosphate-stabilized closed state of the NBDs of the maltose transporter[16], and probably reflects the nucleotide-bound conformation of Wzm–Wzt (Fig. 2b, c and Extended Data Fig. 5).

The transporter's transmembrane domain is formed by two closely interacting Wzm protomers that contact each other through TM1 and TM5 (Fig. 3a). TM5 is capped at the C terminus by a cluster of conserved aromatic residues that pack against the C-terminal end of TM1 of the opposing Wzm protomer. Strikingly, the Wzm protomers enclose a large channel spanning the entire membrane (Fig. 3b). The channel is constricted near the periplasmic exit as well as the Wzm–WztN interface, yet continuously accessible to a 3.5 Å radius probe, thus capable of accommodating a polysaccharide. The structure of the native substrate of *Aa*Wzm–Wzt is currently unknown but a model of the *E. coli* O9a polymannose antigen can be accommodated, with eight to ten sugar units spanning the channel (Fig. 3c).

The channel is lined with aromatic residues that are organized in three layers. First, Tyr18, Trp27 and Trp31 reside at the cytosolic Wzm–Wzm interface where the channel is widest. Second, Tyr39, Phe69, Trp71, Phe72, Phe180 and Trp181 form a central layer halfway across the membrane, and Phe43, Tyr60 and Phe195 surround the periplasmic channel exit (Fig. 3b, d). Protein–carbohydrate interactions are frequently mediated by CH–π stacking interactions between aromatic residues and the sugar rings[17]. Clustering of these residues within the channel strongly suggests a role in O-antigen coordination during transport. Indeed, a continuous and mostly conserved 'aromatic path' runs from the putative cytosolic substrate entrance to the periplasmic channel exit (Fig. 3d and Extended Data Fig. 2). Similar paths have been described in cellulose synthase, cellobiohydrolase 1 and maltoporin[18–20]. As discussed for the maltoporin channel, hydrophobic interactions with aromatics are often combined with a continuous pattern of hydrogen-bond donors and acceptors that contact the hydroxyl groups of the polymers and probably minimize translocation energy barriers[20]. Tyr39, Ser75, Asn76, Ser79, Arg80, Glu110 and Gln177 may serve this purpose in Wzm–WztN (Fig. 3d).

At the periplasmic channel exit, the PG1 helix is also preceded by a conserved aromatic residue, usually a tyrosine (Tyr187), similar to the gate helix on the cytosolic side (Fig. 3c). It is thus likely that PG1 forms the gate towards the periplasmic membrane leaflet. The functional importance of the gate helices was addressed by introducing

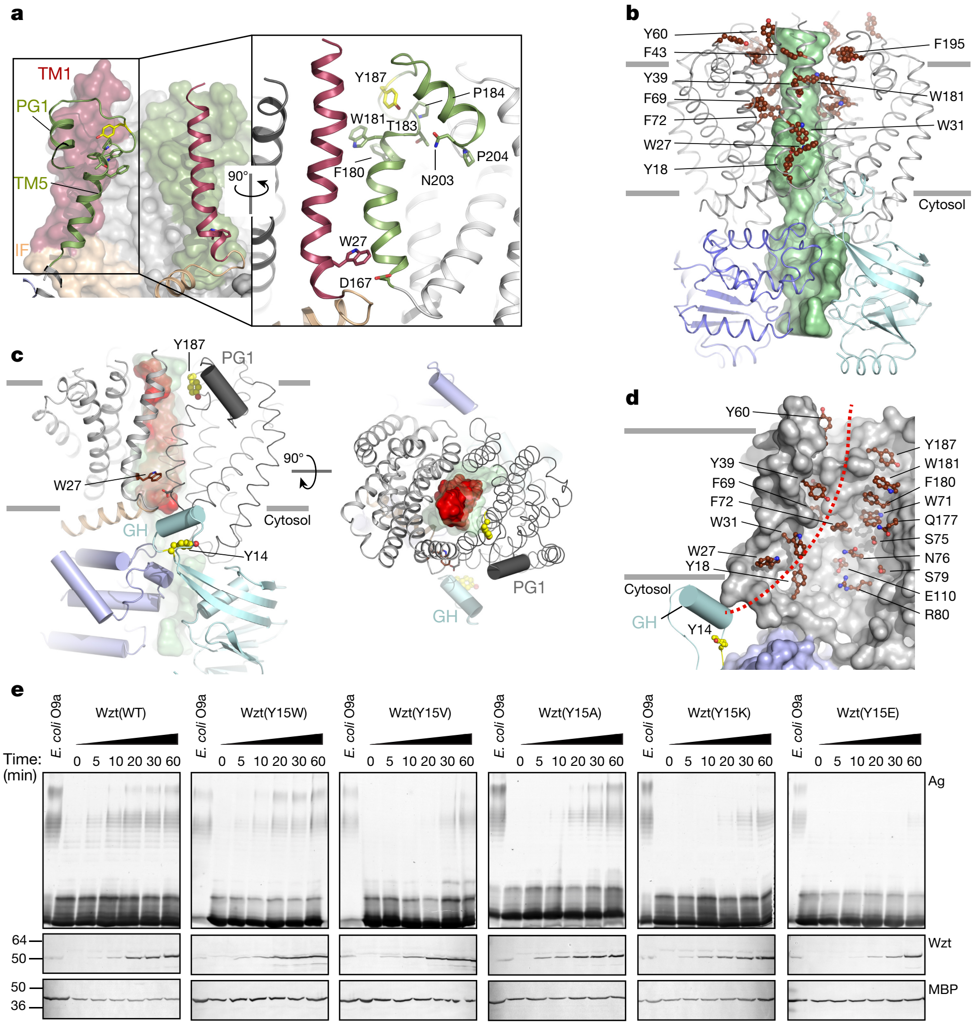


Figure 3 | The polysaccharide translocation channel. a, The Wzm interface. One Wzm protomer is shown as a surface and the opposing subunit is shown as a cartoon. Both subunits are shown as cartoons in the close-up view. TM1 and TM5 are coloured red and green, respectively, and the interface helix is coloured beige. Conserved residues are shown as sticks. **b**, Surface representation of the Wzm–WztN channel. The channel volume accessible to a 3.5 Å-radius probe is shown as a green surface and aromatic residues lining the channel are shown as brown spheres. Selected residues are labelled. **c**, Cytosolic and periplasmic gate helices at the Wzm protomer interface. A model of the *E. coli* O9a antigen containing ten mannose units was manually placed in the channel and is shown as a red surface. **d**, Putative translocation path (red dotted line). Channel-exposed aromatic and polar residues are shown as brown sticks. **e**, *In vivo* O-antigen translocation. The indicated point mutations were introduced into *E. coli* O9a Wzm–Wzt. O-antigen export was detected after inducing transporter expression by silver staining (Ag) of whole-cell lysates, detecting exported and lipopolysaccharide-linked O-antigens only. Western blots detecting Wzt and maltose-binding protein (MBP) were performed to monitor transporter expression and as a loading control, respectively. All results have been confirmed at least three times as technical replicates. Time, period after inducing Wzt–Wzm expression.

point mutations into the *E. coli* Wzm–Wzt O9a transporter and monitoring O-antigen export *in vivo*[8,9]. Tyr14 and 187 of *Aa*Wzm–Wzt correspond to Tyr15 and 192 in *E. coli* Wzm–Wzt, respectively (Extended Data Fig. 2). As shown in Fig. 3e and Extended Data Fig. 6, replacing Tyr15 at the cytosolic gate of Wzt with Trp or Phe supports O-antigen export similar to wild-type levels, whereas replacing it with the hydrophobic β-branched residues Val and Ile abolishes export. Among the charged residues, Lys and Arg support some export, requiring longer incubation periods (post-induction) before reaching detectable levels, whereas the Y15E mutant is inactive. Replacing Tyr15 with Ala or Leu only shows a kinetic effect, the exported O-antigen levels reach wild-type levels about 30 min after initiating transport. This is possibly due to a second Tyr directly N-terminal to the conserved residue, which could functionally replace Tyr15 in *E. coli* Wzt. We were unable to express a Wzt double Tyr mutant to test this hypothesis. These results are consistent with the cytosolic gate forming the substrate-binding pocket, perhaps through CH–π stacking interactions of the tyrosine with the first sugar moiety of the substrate. On the periplasmic side, Tyr192 of Wzm is less critical for export, as most of the tested substitutions supported O-antigen secretion (Extended Data Fig. 6). It is likely that, once initiated, translocation is completed even with a compromised periplasmic gate and/or that other aromatic residues nearby functionally replace Tyr192.

The CBD extending the NBD of the transporter (Extended Data Fig. 2c) interacts with the modified terminus of the O-antigen substrate[8,9]. Crystal structures of isolated CBDs from *Raoultella terrigena* and *E. coli* O9a reveal CBD dimers that are stabilized by intermolecular β-strand exchange[9,10] and, accordingly, the CBD of *Aa*Wzt also purifies as a dimer (Extended Data Fig. 7). Binding of the O-antigen cap probably occurs on the surface of the jelly-roll fold[9,10] (Fig. 4a), but the precise orientation of the CBD relative to the NBD remains unknown.

In the absence of the CBD, Wzm–WztN hydrolyses ATP in a detergent-solubilized state in a tested temperature range from 27 to 65 °C, with an apparent Michaelis constant (K_m) for ATP of about 350 μM at 27 °C (Fig. 4b and Extended Data Fig. 8). Strikingly, the full-length transporter hydrolyses ATP about seven times faster than the truncated

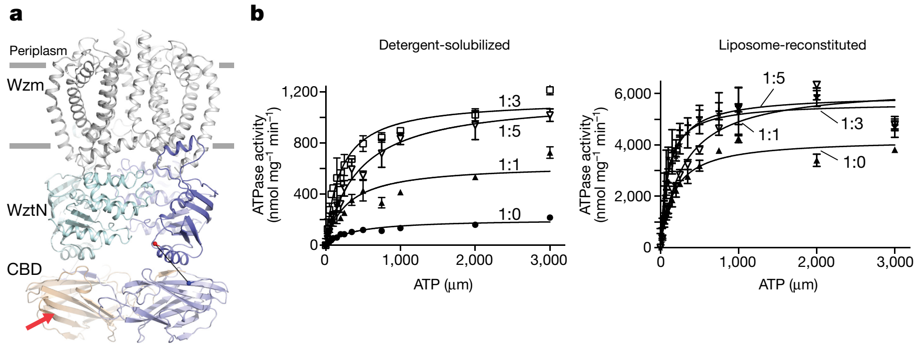


Figure 4 | The CBD stimulates the hydrolytic activity of Wzm–WztN. **a**, Putative organization of the full-length Wzm–Wzt transporter. Alignment of the CBD structure of *E. coli* O9a Wzt (Protein Data Bank accession number 2R5O) with the Wzm–WztN transporter. C and N termini of WztN and CBD are shown as red and blue spheres, respectively. Red arrow, putative binding site of the modified O-antigen cap. **b**, Hydrolytic activity of Wzm–WztN in detergent-solubilized and lipid-reconstituted states, respectively. ATP hydrolysis was performed with increasing CBD concentrations as indicated (Wzm–WztN:CBD, molar ratio). Error bars, s.d. from the means of at least three independent replicates.

version, but a similar apparent K_m for ATP suggests that the CBD accelerates the rate-liming step of ATP hydrolysis (Extended Data Fig. 8c).

To investigate a direct interaction of Wzm–WztN with the CBD, we measured its hydrolytic activity in the presence of increasing CBD concentrations in detergent-solubilized and liposome-reconstituted states. The ATPase activity of Wzm–WztN increases with increasing CBD concentrations and reaches maximum rates at an approximately threefold molar excess of CBD over Wzm–WztN, consistent with a direct CBD–NBD interaction (Fig. 4b). Control experiments in which purified CBD was added to the full-length transporter did not increase its hydrolytic activity. Instead, we observed a slight reduction in ATPase activity, perhaps because of non-specific interactions of the isolated and NBD-attached CBDs (Extended Data Fig. 8b).

Compared with detergent-solubilized states, the hydrolytic activities of the transporters increase significantly upon reconstitution into liposomes (Extended Data Fig. 8c). Assuming similar concentrations of catalytically active transporters, the apparent catalytic rates in liposomes increase about 3- and 20-fold for full-length and truncated Wzm–Wzt, respectively, relative to the detergent-solubilized states. These data suggest that the transporter adopts a different, perhaps closed, conformation in a lipid bilayer environment or the presence of the CBD, thereby affecting its hydrolytic activity. These properties could be modulated by the O-antigen to facilitate translocation.

In the absence of a translocating substrate *in vivo*, the transporter's transmembrane channel must be closed to prevent leakage of small solutes across the membrane. Channel closure probably correlates with closing of the transporter's NBDs, perhaps through rigid body movements of the Wzm–Wzt half transporters relative to one another. This state can be modelled by superimposing the NBDs of the Wzm–WztN transporter halves with the closed structure of the isolated WztN dimer (Fig. 2c). In this model, the transmembrane channel is closed because the Wzm subunits pack tightly against each other without any significant backbone clashes (Extended Data Fig. 9). However, significant overlaps occur at the putative cytosolic substrate-binding site, where the gate helix of WztN contacts the interface helix–TM1 loop of the opposing Wzm subunit (Fig. 2). This region probably undergoes additional conformational changes during channel closure to facilitate substrate translocation (discussed below). The predicted rigid body movement of the transporter halves is supported by disulfide cross-linking of the Wzm subunits. Cys residues introduced into periplasmic loops predicted to be in close proximity in the closed conformation indeed form disulfide bridges under oxidizing conditions (Extended Data Fig. 9c).

The channel-forming conformation of the ABC transporter is consistent with its biological function. However, in the absence of a polysaccharide, mechanisms must exist that prevent spontaneous transporter opening. It is possible that channel formation is tightly coupled to substrate recognition and insertion, such that the translocating polymer seals the channel (Fig. 5).

It is unknown which end of the O-antigen enters the transporter first. Our structural and functional data argue that the gate helix of Wzt functions in substrate binding, most probably by recognizing the pyrophosphate group of UND-PP, together with the first sugar unit. Accordingly, some ABC transporters for O-antigens and teichoic acids that operate without the fine specificity imposed by CBDs can export glycans with different repeat-unit structures[21,22]. Yet, all substrates contain an acetylated amino sugar as the connector between UND-PP and the repeat-unit glycan[7]. Our data suggest that Wzm–Wzt specifically recognizes this motif, in contrast to PglK, which has been proposed to recognize the undecaprenyl moiety[6] (Extended Data Fig. 4b).

Substrate binding to the cytosolic entrance probably leads to opening of the transporter and insertion of the lipid head group into the channel through a gate between the Wzm subunits (Fig. 5). Following insertion, the lipid anchor may spontaneously re-orient to the periplasmic side, possibly facilitated by the proton-motive force across the inner membrane. During this transition, the hydrophobic

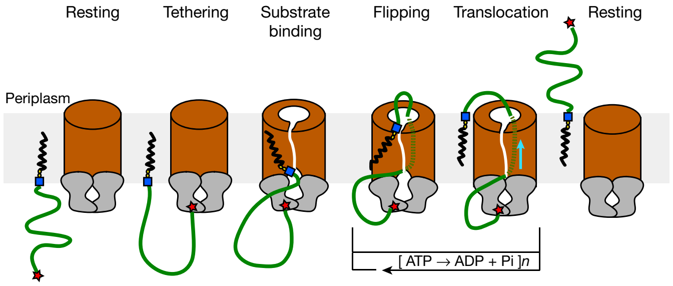


Figure 5 | Model of O-antigen membrane translocation. In a resting state, the transmembrane channel and the NBDs are in a closed conformation. Tethering the substrate to the transporter via interactions of the CBD with the modified O-antigen terminus increases its local concentration. Binding of the UND-PP lipid anchor to the cytosolic gate induces NBD and transmembrane channel opening. The lipid head group inserts into the channel and re-orients spontaneously to the periplasmic side. The now channel-inserted polysaccharide is translocated through repeated cycles of ATP binding and hydrolysis (indicated by *n*). Upon polymer release to the periplasmic side, the transporter returns to the resting conformation with a closed transmembrane channel. Blue square, *N*-acetylglucosamine; yellow spheres, phosphate; red star, modified terminus.

part of the lipid anchor probably remains in the membrane, similar to the model proposed for PglK[6]. After this passive flipping, the transporter contains the polysaccharide in the channel proper. Export could be achieved in a single cycle or require several steps of ATP hydrolysis, however, these alternatives are currently impossible to distinguish.

We speculate that the loop connecting the interface helix of Wzm with TM1 near the cytosolic gate (Figs 1 and 2) contacts the polysaccharide during NBD closure. The gate helix, upon ATP binding, probably pushes against this loop, such that it moves horizontally towards the channel (Fig. 2). The loop contains several polar residues, including Thr21, which could interact with and move the polysaccharide during this transition, similar to the translocation mechanism proposed for cellulose synthase[23]. Conformational changes at the gate could mediate the translocation of about one or two sugar units at a time. As such, Wzm–Wzt would combine the functions of a lipid flippase and polysaccharide translocase.

ABC transporters exporting uncapped O-antigens (for example, *K. pneumoniae* O2a) do not contain C-terminal CBDs[22]. In these systems, polymer export is dependent on simultaneous synthesis, whereas in *E. coli* O9a, O-antigen synthesis and export can be temporally uncoupled[9,22]. Both types of transporter share structural features key to our model (Extended Data Fig. 2). Whereas uncapped O-antigens may be synthesized and exported by multi-subunit complexes including the transporter, CBD-containing ABC transporters probably function independently during or after O-antigen biosynthesis. In this scenario, the CBD may ensure a sufficient local substrate concentration.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 31 July; accepted 1 December 2017.
Published online 10 January 2018.

1. Whitfield, C., Szymanski, C. M. & Aebi, M. in *Essentials of Glycobiology* 3rd edn (ed. A. Varki) Ch. 21, 265–282 (Cold Spring Harbor Laboratory Press, 2017).
2. Raetz, C. R. H. & Whitfield, C. Lipopolysaccharide endotoxins. *Annu. Rev. Biochem.* **71,** 635–700 (2002).
3. Caboni, M. *et al.* An O antigen capsule modulates bacterial pathogenesis in *Shigella sonnei*. *PLoS Pathog.* **11,** e1004749 (2015).
4. Goebel, E. M., Wolfe, D. N., Elder, K., Stibitz, S. & Harvill, E. T. O antigen protects *Bordetella parapertussis* from complement. *Infect. Immun.* **76,** 1774–1780 (2008).
5. Skurnik, M. & Bengoechea, J. A. The biosynthesis and biological role of lipopolysaccharide O-antigens of pathogenic *Yersiniae*. *Carbohydr. Res.* **338,** 2521–2529 (2003).
6. Perez, C. *et al.* Structure and mechanism of an active lipid-linked oligosaccharide flippase. *Nature* **524,** 433–438 (2015).
7. Liston, S. D., Mann, E. & Whitfield, C. Glycolipid substrates for ABC transporters required for the assembly of bacterial cell-envelope and cell-surface glycoconjugates. *Biochim. Biophys. Acta* **1862,** 1394–1403 (2017).
8. Cuthbertson, L., Powers, J. & Whitfield, C. The C-terminal domain of the nucleotide-binding domain protein Wzt determines substrate specificity in the ATP-binding cassette transporter for the lipopolysaccharide O-antigens in *Escherichia coli* serotypes O8 and O9a. *J. Biol. Chem.* **280,** 30310–30319 (2005).
9. Cuthbertson, L., Kimber, M. S. & Whitfield, C. Substrate binding by a bacterial ABC transporter involved in polysaccharide export. *Proc. Natl Acad. Sci. USA* **104,** 19529–19534 (2007).
10. Mann, E., Mallette, E., Clarke, B. R., Kimber, M. S. & Whitfield, C. The *Klebsiella pneumoniae* O12 ATP-binding cassette (ABC) transporter recognizes the terminal residue of its O-antigen polysaccharide substrate. *J. Biol. Chem.* **291,** 9748–9761 (2016).
11. van der Es, D., Hogendorf, W. F., Overkleeft, H. S., van der Marel, G. A. & Codée, J. D. Teichoic acids: synthesis and applications. *Chem. Soc. Rev.* **46,** 1464–1482 (2017).
12. Locher, K. P. Mechanistic diversity in ATP-binding cassette (ABC) transporters. *Nat. Struct. Mol. Biol.* **23,** 487–493 (2016).
13. Lee, J. Y. *et al.* Crystal structure of the human sterol transporter ABCG5/ABCG8. *Nature* **533,** 561–564 (2016).
14. Qian, H. *et al.* Structure of the human lipid exporter ABCA1. *Cell* **169,** 1228–1239 (2017).
15. Hug, I. & Feldman, M. F. Analogies and homologies in lipopolysaccharide and glycoprotein biosynthesis in bacteria. *Glycobiology* **21,** 138–151 (2011).
16. Oldham, M. L. & Chen, J. Snapshots of the maltose transporter during ATP hydrolysis. *Proc. Natl Acad. Sci. USA* **108,** 15152–15156 (2011).
17. Spiwok, V. CH/π interactions in carbohydrate recognition. *Molecules* **22,** http://dx.doi.org/10.3390/molecules22071038 (2017).
18. Morgan, J. L., Strumillo, J. & Zimmer, J. Crystallographic snapshot of cellulose synthesis and membrane translocation. *Nature* **493,** 181–186 (2013).
19. Divne, C., Ståhlberg, J., Teeri, T. T. & Jones, T. A. High-resolution crystal structures reveal how a cellulose chain is bound in the 50 Å long tunnel of cellobiohydrolase I from *Trichoderma reesei*. *J. Mol. Biol.* **275,** 309–325 (1998).
20. Meyer, J. E. & Schulz, G. E. Energy profile of maltooligosaccharide permeation through maltoporin as derived from the structure and from a statistical analysis of saccharide–protein interactions. *Protein Sci.* **6,** 1084–1091 (1997).
21. Schirner, K., Stone, L. K. & Walker, S. ABC transporters required for export of wall teichoic acids do not discriminate between different main chain polymers. *ACS Chem. Biol.* **6,** 407–412 (2011).
22. Kos, V., Cuthbertson, L. & Whitfield, C. The *Klebsiella pneumoniae* O2a antigen defines a second mechanism for O antigen ATP-binding cassette transporters. *J. Biol. Chem.* **284,** 2947–2956 (2009).
23. Morgan, J. L. *et al.* Observing cellulose biosynthesis and membrane translocation *in crystallo*. *Nature* **531,** 329–334 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Morgan and G. Lipkowitz for cloning the full-length Wzm–Wzt transporter, S. Johnson for help with Wzm–WztN purification, C. Caffalette for comments on the manuscript, O. Pornillos for access to SEC-MALS, and the staff at the APS Southeast Regional (SER)- and GM/CA-CAT as well as NSLS-II AMX and FMX beam lines. GM/CA@APS has been funded in whole or in part with federal funds from the National Cancer Institute (ACB-12002) and the National Institute of General Medical Sciences (AGM-12006). The Northeastern Collaborative Access Team beam lines are funded by the National Institute of General Medical Sciences from the National Institutes of Health (NIH) (P41 GM103403). The Pilatus 6M detector on the 24-ID-C beam line is funded by a NIH–ORIP HEI grant (S10 RR029205). Data for this research was also in part collected at the SER Collaborative Access Team beam line at the Advanced Photon Source, Argonne National Laboratory. This research used resources of the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under contract number DE-AC02-06CH11357; and the AMX and FMX beamlines of the National Synchrotron Light Source II, a US DOE Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under contract number DE-SC0012704. Y.B. and J.Z. were supported by NIH grant 5R01GM110143. E.M. is a recipient of an Alexander Graham Bell Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council. C.W. was supported by funding from the Canadian Institutes of Health Research. C.W. is a recipient of a Canada Research Chair.

Author Contributions Y.B. and J.Z. designed all structural and *in vitro* biochemical experiments, and E.M and C.W. designed the *in vivo* functional assays. Y.B. and E.M. performed all structural and *in vivo* functional experiments, respectively. Y.B. and J.Z. wrote, and all authors edited, the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.Z. (jochen_zimmer@virginia.edu).

Reviewer Information *Nature* thanks R. Ford and W. Vollmer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Cloning and protein expression. The *wzm* and *wzt* genes were PCR amplified from genomic *A. aeolicus* VF5 DNA and sequentially cloned into an engineered pETDuet expression vector (Novagen) with a C-terminal histidine tag on Wzt. A second construct containing only residues 1–235 of Wzt (WztN) was cloned in a similar manner. The transporters were expressed in *E. coli* C43 cells in Luria broth (LB) medium upon induction with 0.5 mM isopropyl-β-D-thiogalactoside (IPTG) at an absorbance at 600 nm of 0.6. Cells were harvested by centrifugation after incubation at 37 °C for 4 h. The cells were resuspended in RB-1 buffer containing 20 mM Tris HCl pH 7.5, 0.1 M NaCl, and 5 mM β-mercaptoethanol (β-ME) and then lysed in a microfluidizer. The crude membranes were collected by centrifugation for 60 min at 200,000*g* in a Beckman Ti45 rotor and solubilized for 60 min at 4 °C in RB-2 buffer containing 50 mM sodium phosphate pH 7.2, 0.1 M NaCl, 20 mM imidazole, 5 mM β-ME, and 2% polyoxyethylene(8)-dodecyl ether (C12E8). The insoluble material was cleared by centrifugation for 30 min at 200,000*g* in a Beckman Ti45 rotor and the membrane extract was batch incubated with Ni-NTA agarose (Qiagen) for 60 min at 4 °C. The resin was packed in a gravity flow chromatography column, washed with 50 ml WB1 buffer (RB-1 buffer containing 20 mM imidazole and 5 mM dodecyl-*N*,*N*-dimethylamine-*N*-oxide (LDAO)), 50 ml WB2 buffer (RB-1 buffer containing 40 mM imidazole and 5 mM LDAO), and 50 ml WB3 buffer (RB-1 buffer containing a total of 1.5 M NaCl, 20 mM imidazole and 5 mM LDAO), and the transporter was eluted in 50 ml EB buffer containing 20 mM Tris HCl pH 7.5, 0.1 M NaCl, 300 mM imidazole, 5 mM β-ME, and 5 mM LDAO. The eluted protein was purified over an analytical S200 gel filtration column (GE Healthcare) equilibrated in buffer containing 20 mM Tris HCl pH 7.5, 0.1 M NaCl, 0.5 mM tris(2-carboxyethyl)phosphine (TCEP), and 5 mM LDAO. The peak fraction was concentrated to 15 mg ml^{-1} final concentration in a 50-kDa cut-off centrifugal filter (Amicon) before crystallization in the presence of 5 mM $MgCl_2$. To guide model building, Thr128 in TM3 of Wzm was replaced with a Cys residue for derivatization with ethylmercurithiosalicylic acid. This mutant was generated by QuikChange mutagenesis and purified as described for the wild-type transporter. Selenomethionine-derivatized Wzm–WztN was prepared as described above with the exception that the cells were grown in the M9 minimal medium supplemented with 60 μg ml^{-1} L-selenomethionine (Se-Met).

The CBD of Wzt (residues 235–394) was expressed from a pET30a vector (Novagen) in *E. coli* BL21 (DE3) cells (Invitrogen). The *E. coli* cells were cultured in LB medium at 37 °C and protein expression was induced at an optical density at 600 nm of 0.6 with 0.5 mM IPTG. The cells were harvested by centrifugation after 4 h of incubation at 37 °C. Subsequently, the cells were resuspended in RB-1 buffer containing 20 mM sodium phosphate pH 7.5, 0.05 M NaCl and 5 mM β-ME and then lysed in a microfluidizer. The insoluble material was cleared by centrifugation for 30 min at 200,000*g* in a Beckman Ti45 rotor. The supernatant was batch incubated with Ni-NTA agarose (Qiagen) for 60 min at 4 °C. The protein was then purified by Ni-NTA affinity chromatography at 4 °C, washed with 50 ml WB1 buffer containing 25 mM Tris HCl pH 8.5, 0.5 M NaCl, 30 mM imidazole, and 5 mM β-ME, 50 ml WB2 buffer (WB1 buffer containing a total of 50 mM imidazole and 50 mM NaCl) and eluted with EB buffer, consisting of 25 mM Tris HCl pH 8.5, 50 mM NaCl, 300 mM imidazole, and 5 mM β-ME. The protein was further purified by gel filtration chromatography (Superdex-200) in 25 mM Tris, pH 8.5, 50 mM NaCl, 5 mM β-ME.

Crystallization. The truncated Wzm–WztN transporter was crystallized by combining 1 μl of well solution (32% polyethylene glycol (PEG) 400, 0.05 M sodium acetate pH 5.4, and 0.1 M magnesium acetate) with 1 μl of protein solution and sitting-drop vapour diffusion at 22 °C. Addition of 3.6 mM decyl-β-D-maltoside to the crystallization solution significantly improved diffraction. Crystallization trials with full-length Wzm–Wzt were unsuccessful.

WztN crystallized from a Wzm–WztN sample set up under different conditions, each producing a $P3_121$ crystal form but with different unit cell dimensions. Crystals with a smaller unit cell contained a WztN monomer in the crystallographic asymmetric unit and were obtained by sitting-drop vapour diffusion in the presence of 41% PEG 400, 0.05 M sodium acetate pH 5.4, 0.15 M magnesium acetate, and 47 mM octyl-glucoside at 17 °C. Crystals with a larger unit cell contained a WztN dimer per asymmetric unit and grew in the presence of 0.4 M magnesium nitrate, 17.5% PEG 8000, and 0.1 M Tris pH 8.5 by sitting-drop vapour diffusion at 17 °C.

Initial crystals were observed after approximately three days and reached their final size within two weeks for all samples. The crystals were collected and directly cryo-cooled in liquid nitrogen. WztN crystals grown in PEG 8000 were cryoprotected in the presence of 25% glycerol in the crystallization solution. Wild-type and Wzm-T128C crystals were soaked with 5–20 mM ethylmercurithiosalicylic acid for 2–20 h before harvesting.

Structure determination. Diffraction data were collected at the Argonne National Laboratory beam lines SER- and GM/CA-CAT as well as the AMX and FMX beam lines at the Brookhaven National Laboratories (NSLS-II). Data were integrated in XDS and reduced in Aimless, as part of the CCP4 program suite[24]. The isolated WztN structure was determined after single anomalous dispersion phasing with an ethylmercurithiosalicylic-acid-derivatized crystal containing a WztN monomer per asymmetric unit in Phenix.autosol with subsequent model building[25,26]. The obtained model was of sufficient quality for molecular replacement using Phaser with the native data set[27], after which the model was manually completed in Coot[28] and refined in Phenix.refine[25] using TLS parameters[29]. In this crystal form, the protomers of a WztN dimer are related by crystallographic two-fold symmetry. The obtained model was used for molecular replacement of the crystal form containing a WztN dimer per asymmetric unit and the obtained model was completed in Coot and refined as described above. In this structure, residues 12–33 of chain A were disordered and not included in the final model. The final monomeric and dimeric WztN models contained 99.2/0.4/0.4% and 98.4/1.6/0% in the preferred, allowed, and outlier regions of the Ramachandran plot.

The *Aa*Wzm–WztN structure was determined by single anomalous dispersion phasing with an ethylmercurithiosalicylic-acid-derivatized wild-type Wzm–WztN crystal. Phasing was performed in Phenix.autosol without automated model building and based on five mercury sites near Cys residues in the transporter's NBDs. The obtained experimental phases to approximately 9 Å resolution were improved by non-crystallographic symmetry (NCS) averaging and solvent flattening using the program DM with manually built averaging and solvent masks[30,31]. The improved phases allowed manual docking of the WztN NBDs as well as three partial transmembrane helices. The obtained poly-alanine model was used to improve averaging and solvent masks for subsequent rounds of NCS averaging. Following placement of another two partial transmembrane helices, the initial model was of sufficient quality for molecular replacement with the native and highest-resolution data set. Subsequently, NCS and cross-crystal averaging in DMMulti[31] was used to improve and gradually extend the phases to 3.85 Å resolution using a solvent mask that covered the entire transporter and an averaging mask covering a Wzm–WztN half-transporter. The greatly improved density maps were used to place additional regions as poly-alanine traces, refine the averaging and solvent masks, and iterative rounds of NCS and cross-crystal averaging until the backbone of the entire transporter could be traced and bulky residues were discernible. We confirmed the annotation of transmembrane helices and the assigned registers on the basis of seleno-methionines at positions 147 and 156 in transmembrane helix 4 and 205 in PG2, as well as a mercury-derivatized Cys residue introduced at position 128 at the C-terminal end of transmembrane helix 3. The initial model was built as a poly-alanine model and refined in Phenix.refine with NCS constraints. To generate a model containing all amino-acid side chains, the high-resolution structure of the NBD was docked into the Wzm–WztN density and manually refined. The Wzm subunit was manually built starting with aromatic amino acids and methionines, and cycles of NCS refinement in Phenix.refine. Additionally, refinement in Refmac5 with jelly-body refinement greatly improved the model[32]. Towards the end of the refinement, NCS restraints were used instead of constraints. The final Wzm–WztN model contains residues 2–235 of Wzt plus a KLHH sequence corresponding to an engineered HindIII restriction site and C-terminal His-tag. Wzm contains residues 2–255 (of 256 in total) with a short gap between residues 51 and 55 in chain C and 49 and 55 in chain D. The final model contains 93.5, 6.3, and 0.2% of the residues in favoured, allowed, and disallowed regions of the Ramachandran plot, respectively. The coordinates and structure factors have been deposited in the Protein Data Bank. All figures were generated using Pymol[33] and channel dimensions were analysed using HOLLOW[34]. Primary sequence alignments were generated in CLUSTALW Omega and displayed in Jalview[35].

ATPase activity measurements. The ATPase activity was measured using an NADH-consuming coupled method as described[36]. In a first step, the transporter was incubated in reaction buffer containing 10 mM $MgCl_2$ and the indicated amounts of ATP at 27 °C (or otherwise indicated temperatures) for 10 min, after which the samples were snap frozen in liquid nitrogen until all data points had been collected. In a second step, the thawed samples were incubated with NADH buffer containing 50 mM HEPES pH7.5, 8 U pyruvate kinase, 8 U lactate dehydrogenase, 4 mM phosphoenolpyruvate, 1 mM $MgCl_2$, and 0.05 mM NADH at 22 °C for 5 min in the dark. At 22 °C, *Aa*Wzm–Wzt does not exhibit any detectable ATPase activity. ADP formation was quantified by following the decrease in NADH fluorescence at 450 nm (excitation at 340 nm) using a FluoroMax 3 (Horiba) fluorimeter at 22 °C. For ATPase assays in proteoliposomes, purified Wzm–Wzt was reconstituted into 3 mg ml^{-1} *E. coli* total lipid extract at a protein concentration of 0.3 mg ml^{-1}. Vesicles were formed upon detergent removal using SM-2 Bio-beads (Bio-rad). For CBD titration experiments, the CBD of Wzt was premixed with transporter for 3 h at 4 °C before adding ATP and $MgCl_2$. Decrease in NADH fluorescence was converted to molar concentrations on the basis of measurements of known standards. All experiments were repeated at least three times and data were fitted to Michaelis–Menten kinetics to calculate K_m and V_{max} values using GraphPad Prism 6. Error bars are deviations from the means.

Disulfide cross-linking of Wzm–WztN. For disulfide cross-linking experiments, Wzm and WztN were co-expressed from pETDuet and pACYC vectors, respectively, with an N-terminal Flag-tag on Wzm and C-terminal His-tag on WztN. Protein purification was as described above with the exception that the gel filtration buffer contained 1 mM DTT instead of TCEP. Disulfide cross-linking experiments confirming the modelled closed conformation of Wzm–WztN were performed with purified detergent-solubilized transporter. Oxidation was induced with copper-phenanthroline or sodium tetrathionate (STT). A copper-phenanthroline stock solution was prepared by combining 0.36 M 1,10-phenanthroline monohydrate (VWR) (in 50% ethanol) with 0.24 M copper sulfate (Sigma) at a 2:1 volume ratio. STT was dissolved in double-distilled H_2O at 80 mM concentration, and *N*-ethylmaleimide to block free cysteines was prepared at 1 M concentration in dimethylsulfoxide. Wild-type or Cys-introduced Wzm–WztN at 0.09 mM concentration was incubated with 4 mM copper-phenanthroline or STT and incubated for 40 min at room temperature followed by addition of 25 mM *N*-ethylmaleimide and incubation at 4 °C for 30 min. Samples oxidized in the presence of ADP/Mg^{2+} were pre-incubated with 2 mM ADP and 2 mM $MgCl_2$ for 20 min at room temperature. The oxidized protein was resolved by non-reducing SDS–PAGE and protein bands were visualized by western blotting against an N-terminal Flag-tag on Wzm.

Size-exclusion multi-angle light scattering. Mass measurements of the CBD of Wzt were performed on a Dionex UltiMate3000 HPLC system with a UV detection module (ThermoFisher), connected to a miniDAWN TREOS static light-scattering detector (Wyatt Technology) and Optilab T-rEX differential refractometer (Wyatt Technology). A 100-μl sample at 0.1 mM concentration was loaded onto a Superdex 200 HR 10/300 GL column (GE Healthcare) equilibrated in 25 mM Tris, pH 8.5, 50 mM NaCl, 5 mM β-ME at a flow rate of 0.4 ml min^{-1}. Data were recorded and processed using ASTRA software (Wyatt Technology).

***E. coli* O9 antigen modelling.** An *E. coli* O9a antigen containing ten mannose units was modelled using the GLYCAM carbohydrate builder and manually placed into the *Aa*Wzm–WztN transmembrane channel (http://glycam.org/tools/molecular-dynamics/oligosaccharide-builder/build-glycan?id=1).

***In vivo* O-antigen export assays.** *Growth conditions*. Bacterial cultures (Supplementary Information) were grown with aeration in LB base (Invitrogen) at 37 °C. Broth was supplemented with 100 mg ml^{-1} ampicillin, 0.4% D-mannose, and/or 0.1% L-arabinose where appropriate. Unless otherwise stated, cells were grown in the presence of 0.4% D-glucose to repress mannose uptake.

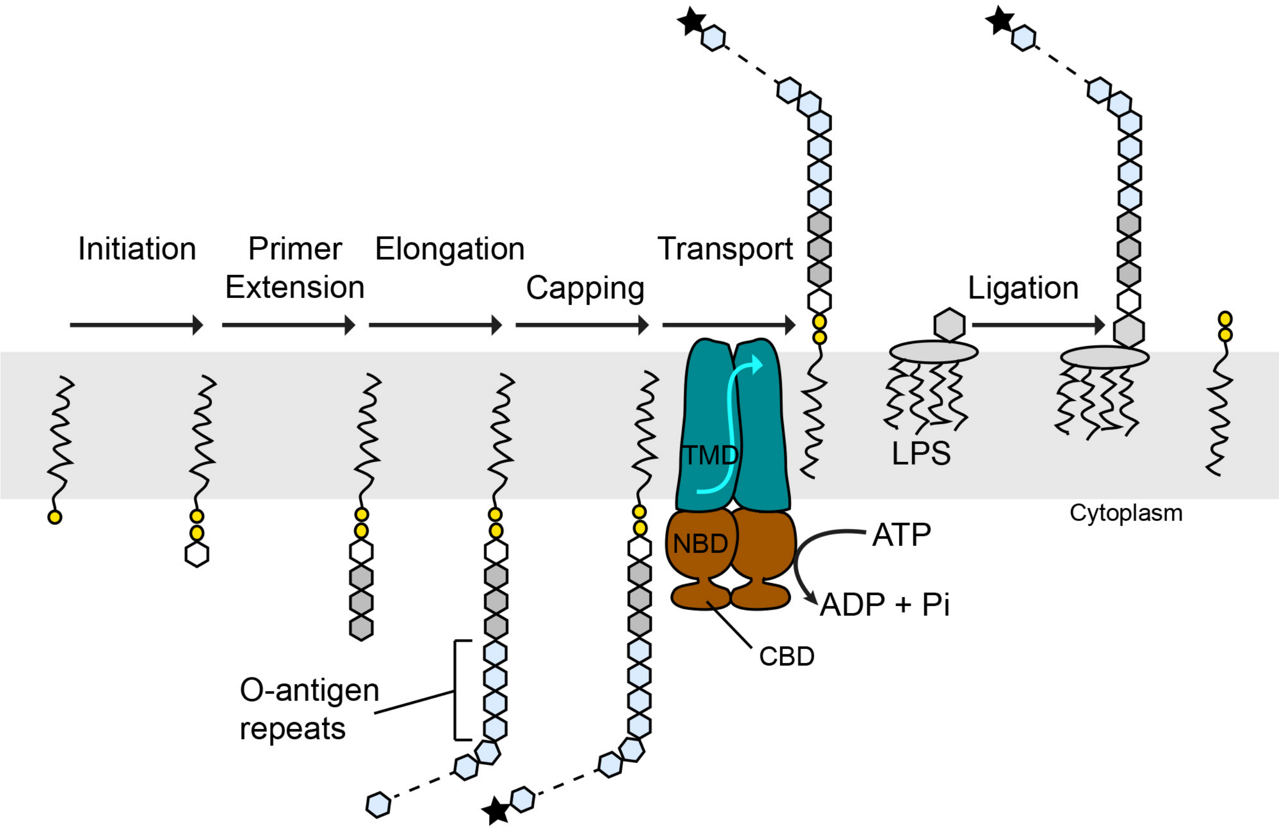
DNA methods. Oligonucleotide primers were custom designed and obtained from Sigma Aldrich (Supplementary Information). PfuUltra DNA polymerase (Agilent) was used for PCR reactions, according to the manufacturer's instructions, and PCR product was treated with DpnI (New England Biolabs). DNA sequencing was performed by the Genomics Facility at the University of Guelph Advanced Analysis Center.

Complementation. *E. coli* CWG638 transformants containing plasmids with wild-type or variant *wzt*, along with *wzm* in the native chromosomal organization, were used to ensure equal protein expression levels. Cultures were grown overnight in the presence of 0.4% D-glucose. *E. coli* CWG638 cannot produce its own GDP-mannose, the substrate of glycosyltransferases responsible for O9a O-antigen assembly, owing to a deletion of *manA*[8]; therefore, O-antigen production relies upon mannose uptake. Accumulation of UND-PP-O9a intermediates in the absence of export results in growth defects that are alleviated by second-site mutations that repress O-antigen synthesis. Growth in glucose represses uptake of any trace amounts of mannose in the medium and prevents harmful O9a synthesis. Overnight cultures were diluted 1/10 in fresh LB supplemented with 0.4% glucose and grown for 4 h to an absorbance at 600 nm of approximately 1.0. Cells were subjected to centrifugation at 5,000*g* for 10 min, resuspended in fresh LB containing 0.4% D-mannose to induce O-antigen biosynthesis, and grown for 15 min at 37 °C with aeration. After 15 min, 0.1% L-arabinose was added to induce protein expression. Aliquots of culture were taken immediately and after 5, 10, 20, 30, and 60 min and put immediately on ice to suppress further cell growth and O-antigen export. An equivalent of 1 optical density cells were harvested by centrifugation at 13,000*g* and resuspended in 100 μl of SDS–PAGE loading buffer. Cells were lysed by boiling for 10 min. For western immunoblots, samples were subjected to SDS–PAGE using 12% acrylamide resolving gels in Tris-glycine buffer[37,38]. For lipopolysaccharide analysis, samples were first treated with 500 μg ml^{-1} proteinase K for 1 h at 55 °C before SDS–PAGE. Lipopolysaccharide was visualized by silver staining[39].

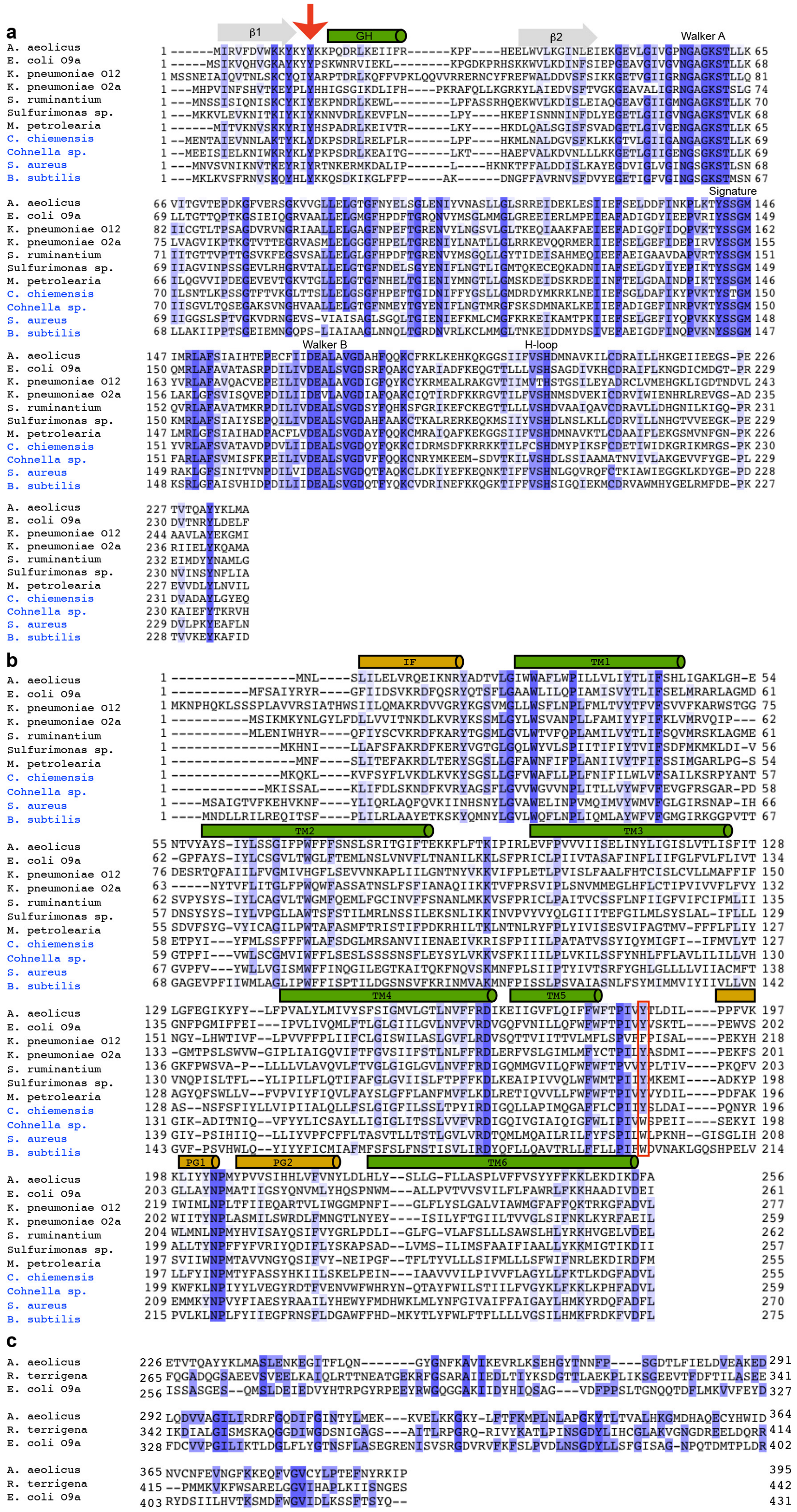
For immunoblot analyses, material resolved by SDS–PAGE was transferred to nitrocellulose membranes (Protran; PerkinElmer Life Sciences). Wzt was detected using anti-Wzt primary antisera, generated in rabbits, and cross-reactive material adsorbed against *E. coli* CWG708 whole-cell lysate[8]. Goat-anti-rabbit secondary antibody conjugated to alkaline phosphatase (Cedarlane Laboratories) was used to facilitate detection with nitroblue tetrazolium and 5-bromo-4-chloro-3-indoyl phosphate (Roche Applied Science). MBP was detected using monoclonal anti-MBP mouse primary antibody (New England Biolabs) with secondary alkaline phosphatase-conjugated goat-anti-mouse antibody (Jackson ImmunoResearch Laboratories) for detection with 5-bromo-4-chloro-3-indoyl phosphate and nitroblue tetrazolium.

Data availability. Atomic coordinates for the atomic models have been deposited in the Protein Data Bank under accession numbers 6AN7 for Wzm–WztN, and 6AMX and 6AN5 for WztN.

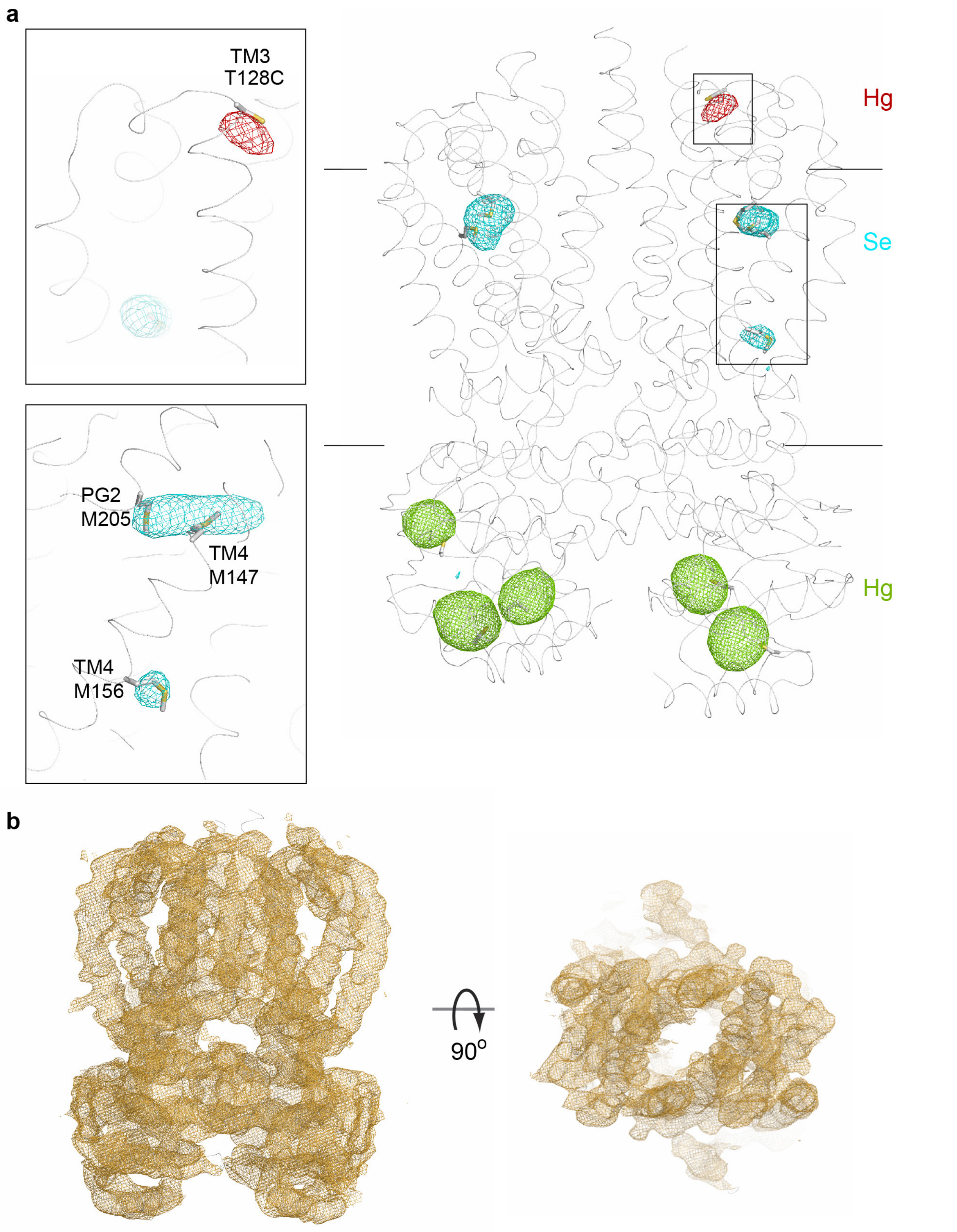
24. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
25. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
26. Zwart, P. H. *et al.* Automated structure solution with the PHENIX suite. *Methods Mol. Biol*. **426,** 419–435 (2008).
27. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
28. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
29. Painter, J. & Merritt, E. A. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr. D* **62,** 439–450 (2006).
30. Cowtan, K. D. & Zhang, K. Y. Density modification for macromolecular phase improvement. *Prog. Biophys. Mol. Biol.* **72,** 245–270 (1999).
31. Cowtan, K. Recent developments in classical density modification. *Acta Crystallogr. D* **66,** 470–478 (2010).
32. Nicholls, R. A., Long, F. & Murshudov, G. N. Low-resolution refinement tools in REFMAC5. *Acta Crystallogr. D* **68,** 404–417 (2012).
33. DeLano, W. L. The PyMOL Molecular Graphics System (2002).
34. Ho, B. K. & Gruswitz, F. HOLLOW: generating accurate representations of channel and interior surfaces in molecular structures. *BMC Struct. Biol.* **8,** 49 (2008).
35. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–1191 (2009).
36. Lin, D. Y., Huang, S. & Chen, J. Crystal structures of a polypeptide processing and secretion transporter. *Nature* **523,** 425–430 (2015).
37. Hitchcock, P. J. & Brown, T. M. Morphological heterogeneity among *Salmonella* lipopolysaccharide chemotypes in silver-stained polyacrylamide gels. *J. Bacteriol.* **154,** 269–277 (1983).
38. Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227,** 680–685 (1970).
39. Tsai, C. M. & Frasch, C. E. A sensitive silver stain for detecting lipopolysaccharides in polyacrylamide gels. *Anal. Biochem.* **119,** 115–119 (1982).
40. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336,** 1030–1033 (2012).



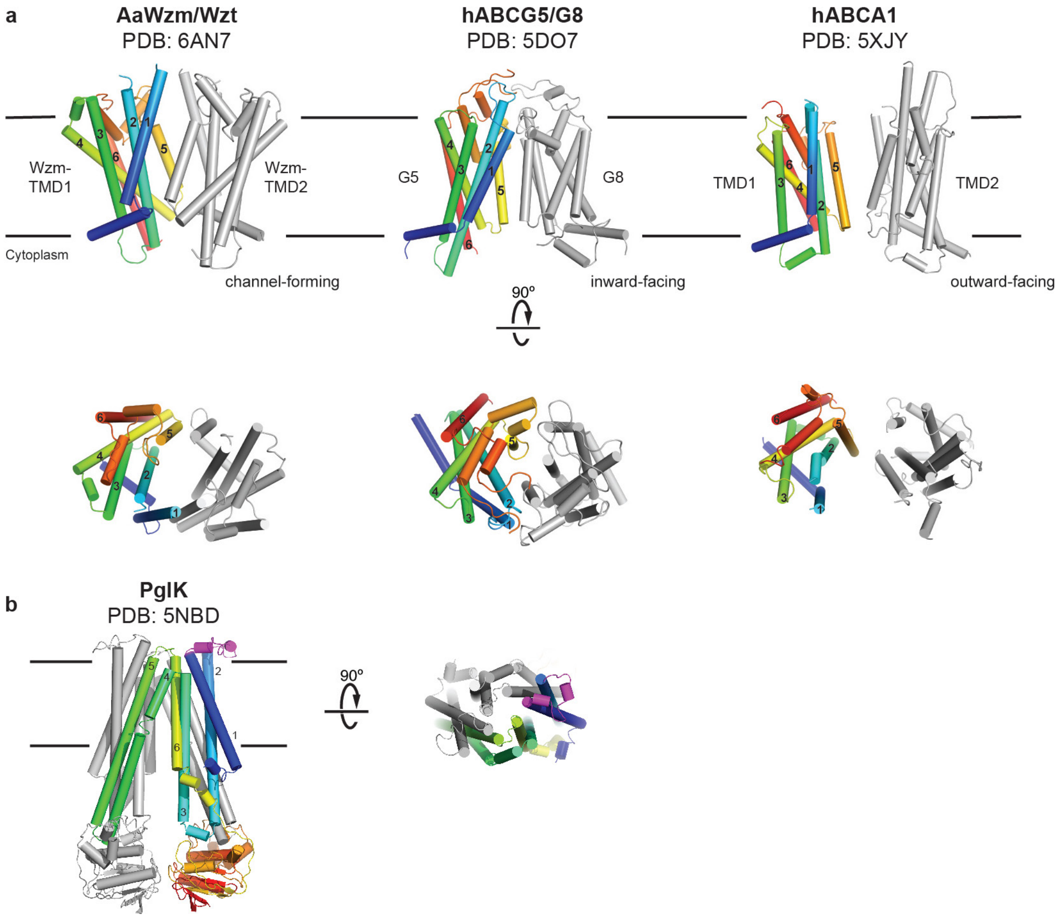
Extended Data Figure 1 | ABC transporter-dependent O-antigen biosynthesis. In this pathway, O-antigens are completely synthesized on the cytosolic leaflet of the plasma membrane. Undecaprenyl-phosphate (black line and yellow circle) serves as the lipid acceptor and is modified by the addition of an acetylated amino sugar phosphate (frequently *N*-acetylglucosamine-1-P, white hexagon) as well as two or more additional sugar residues (grey hexagons) to generate a biosynthesis primer. The polymerizing enzyme(s) extend the primer with tens to hundreds of O-antigen repeat units (light blue hexagons). In some species, termination of O-antigen biosynthesis is achieved by modifying the polymer's non-reducing end (black star). An ABC transporter translocates the UND-PP-linked O-antigen intermediate to the membrane's periplasmic side, where it forms a substrate for glycosylation of the lipopolysaccharide (LPS) core. Only transporters translocating terminally modified O-antigens contain CBDs that bind the polysaccharide's modified terminus.



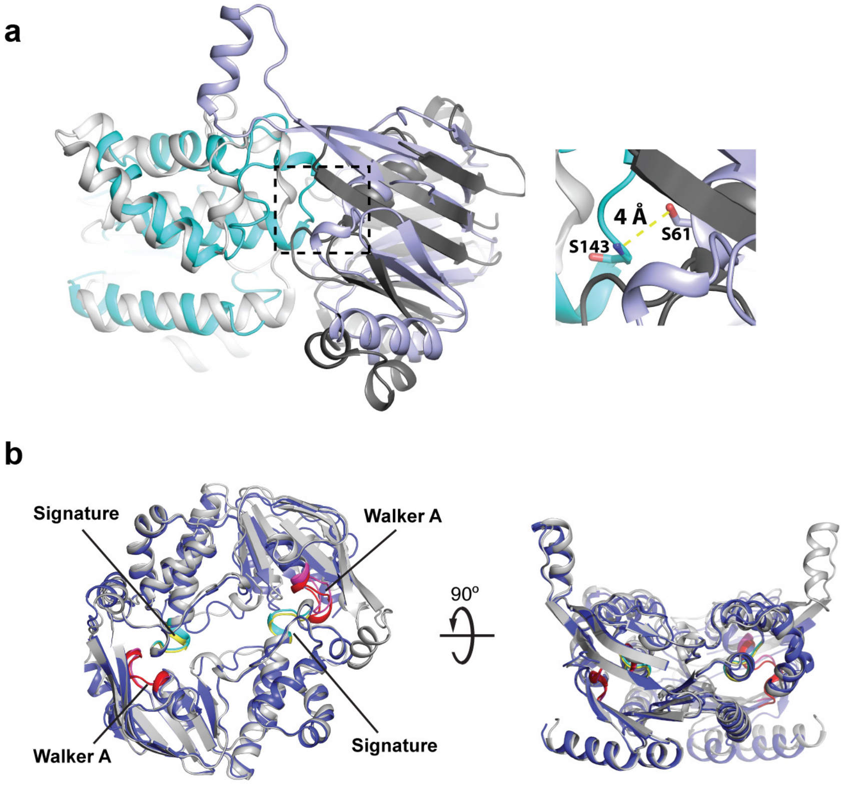
Extended Data Figure 2 | Sequence alignment of O-antigen and wall teichoic acid transporters. a, b, Alignments of the nucleotide-binding (**a**) and transmembrane domains (**b**). The conserved tyrosines preceding the cytosolic gate helix of the NBD and the periplasmic gate are highlighted with a red arrow and red box in **a** and **b**, respectively. Transmembrane helices and cytosolic and periplasmic gate helices are shown as green and beige cylinders, respectively. Blue sequence labels indicate predicted teichoic acid transporters. All O-antigen transporter NBDs except for *K. pneumoniae* O2a contain predicted CBDs at their C termini, which are not shown. **c**, Alignment of the C-terminal region of *Aa*Wzt with the corresponding domains from the *E. coli* O9a (Protein Data Bank accession number 2R5O) and *R. terrigena* (Protein Data Bank accession number 5HNO) transporters. Sequences were aligned in CLUSTAL Omega and displayed in Jalview coloured by sequence identity.



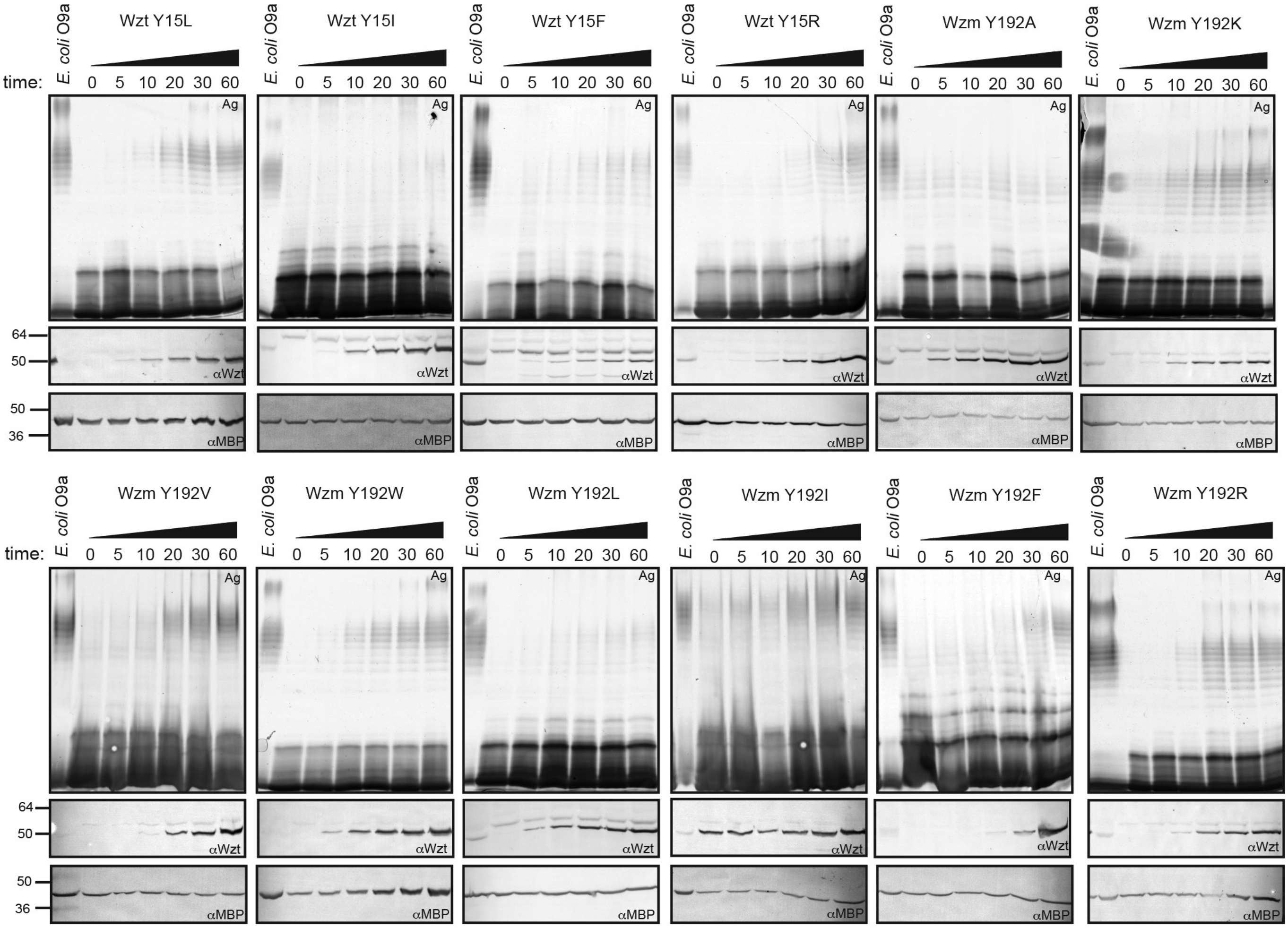
Extended Data Figure 3 | Anomalous difference and experimental electron density maps. a, Heavy atom positions used for experimental phasing and model building. Five native cysteines in the NBDs as well as an engineered Cys at the C terminus of TM3 (T128C) were modified with ethylmercurithiosalicylic acid, shown as green and red meshes and contoured at 4.5σ and 3σ, respectively. Only the mercury sites shown in green were used for mercury-single anomalous dispersion phasing. The transmembrane domain contains three native Met residues, which were identified upon substitution with seleno-methionine (cyan mesh, contoured at 3σ). Shown are sigma-A-weighted anomalous difference electron densities; *Aa*Wzm–WztN is shown as a grey ribbon. **b**, Unbiased experimental sigma-A-weighted electron density after NCS and cross-crystal averaging and phase extension to 3.85 Å, contoured at 1σ.



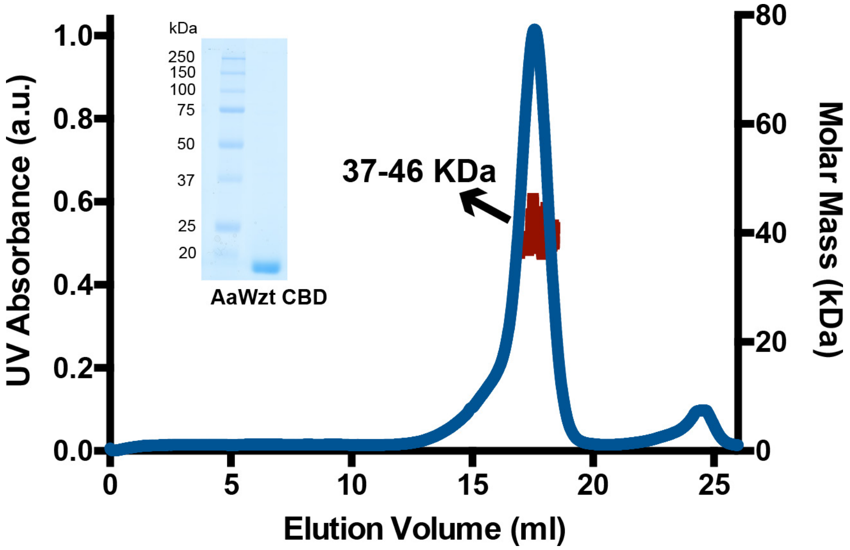
Extended Data Figure 4 | Overview of type-II ABC exporters and PglK. **a**, The structures of the transmembrane domains of *A. aeolicus* Wzm, *Homo sapiens* ABCG5 and *H. sapiens* ABCA1 are shown as cylindrical cartoons. One subunit of the dimers is coloured in rainbow colours from blue to red, N terminus to C terminus. **b**, Structure of PglK, an ABC transporter translocating UND-PP-linked oligosaccharides across the plasma membrane. PglK probably recognizes the polyprenyl moiety of the substrate via a conserved periplasmic helix (shown in magenta), which is missing in Wzm.



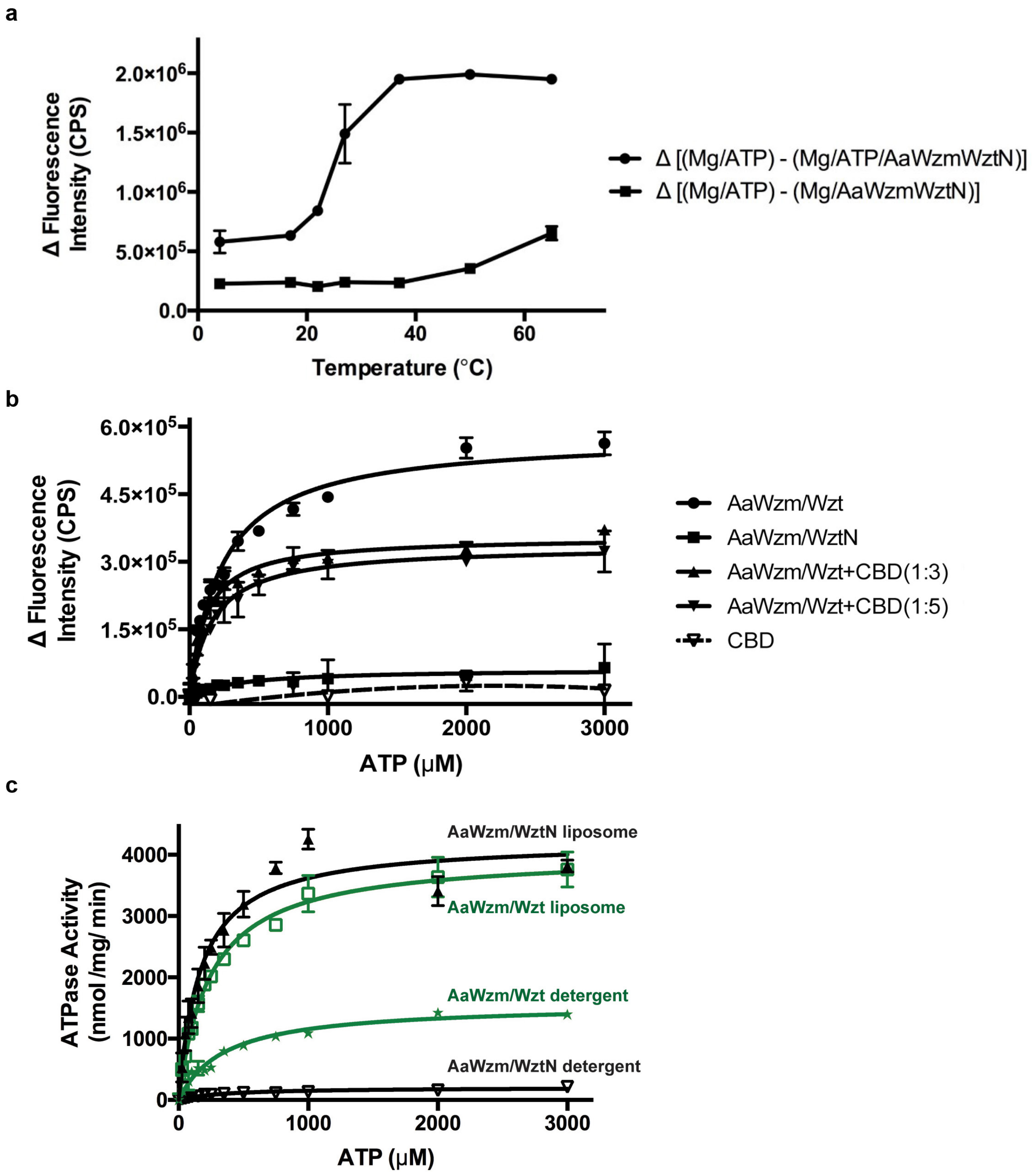
Extended Data Figure 5 | Closed conformation of the isolated WztN-NBD. **a**, The isolated WztN dimer structure was aligned by secondary matching in Coot with the NBDs of the adenosine 5′-(β,γ-imido)triphosphate-stabilized maltose transporter (Protein Data Bank accession number 3RLF). The WztN dimer is shown in cyan and light blue, and the NBDs of the maltose transporter are shown in light and dark grey. Right: the Walker A (S61) and signature (S143) motifs in the closed WztN dimer structure are separated by approximately 4 Å. **b**, Comparison of WztN dimer structures. The structure shown in dark blue was obtained from a crystal form containing a WztN dimer in the crystallographic asymmetric unit. The structure shown in grey was obtained from a crystal form with a monomeric WztN per crystallographic asymmetric unit related to the other protomer by two-fold crystallographic symmetry. The signature motifs are coloured cyan and yellow, and the Walker A motifs are coloured magenta and red for the crystallographic monomeric and dimeric WztN structures, respectively.



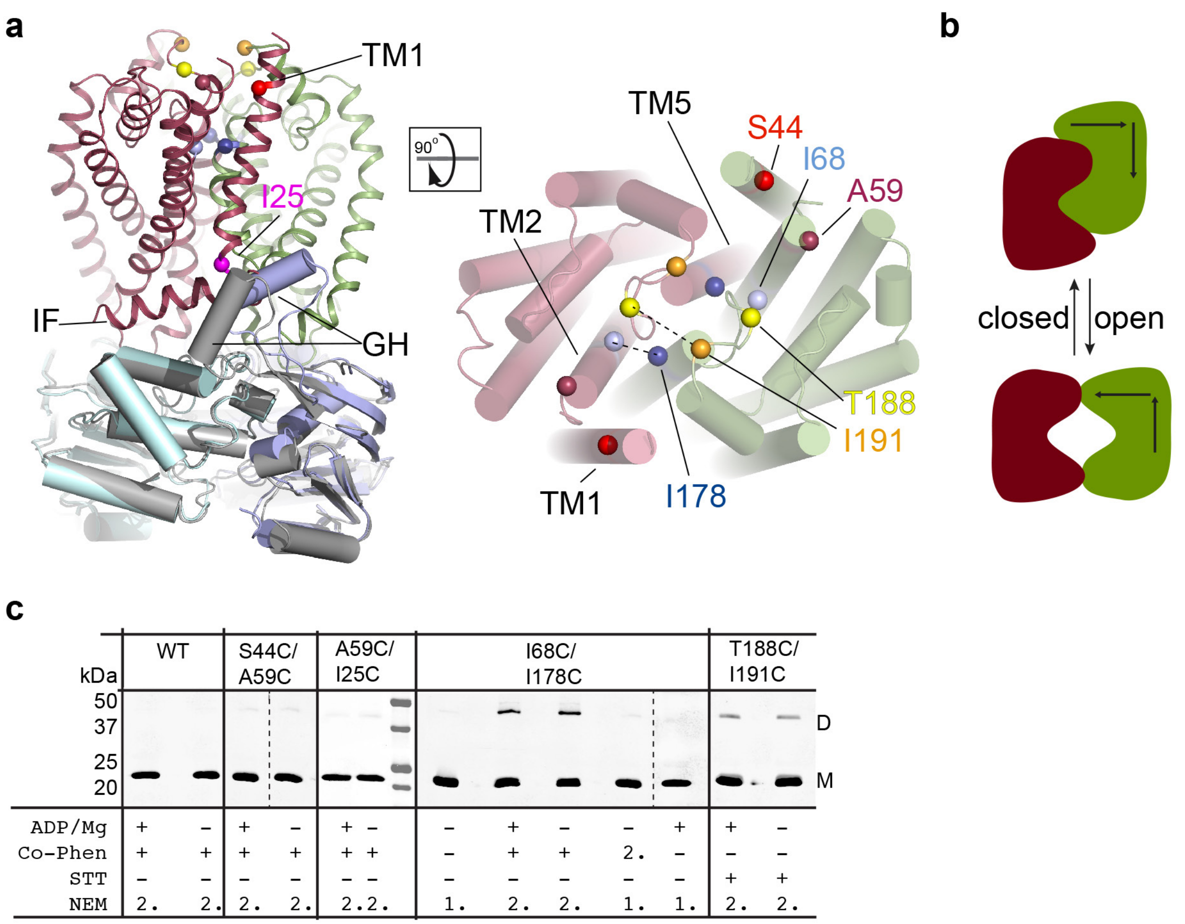
Extended Data Figure 6 | Impact of conserved tyrosine residues of the cytosolic and periplasmic gates on O-antigen translocation. The indicated point mutations were introduced into the *E. coli* O9a Wzt–Wzm transporter and O-antigen transport was assayed by silver staining of the whole-cell lysate. Ag, silver-stained SDS–PAGE. Wzt and MBP were detected immunologically to monitor transporter expression and as a loading control, respectively. All results showing a phenotype have been confirmed at least three times as technical replicates. Time, period after inducing Wzt–Wzm expression in minutes.



Extended Data Figure 7 | Dimerization of the isolated CBD of Wzt. Multi-angle static light scattering coupled to size-exclusion chromatography was used to determine the molecular mass of the purified CBD of Wzt (one representative experiment is shown). The molecular mass of a monomeric Wzt-CBD is 20 kDa, including a C-terminal 6×His-tag and linker region. Inset, Coomassie-stained SDS–PAGE of the purified CBD of Wzt.



Extended Data Figure 8 | Hydrolytic activity of the Wzm–Wzt ABC transporter. ATP hydrolytic activity was measured by following the decrease of NADH fluorescence in an enzyme-coupled assay upon excitation at 340 nm and emission at 450 nm in a temperature range from 4 to 65 °C. **a**, Temperature dependence of the ATPase activity of Wzm–WztN. Shown is the difference in NADH fluorescence between control reactions in the absence of Wzm–WztN and reactions in its presence. **b**, Hydrolytic activity of full-length Wzm–Wzt in the presence of isolated Wzt-CBD measured at 27 °C. Shown are fluorescence intensity differences (calculated as for Fig. 4b) but not converted to apparent catalytic rates. Dashed line, ATP titration in the presence of only the CBD of Wzt. Hydrolytic activity of Wzm–WztN in the absence of the CBD of Wzt is shown for comparison. **c**, Comparison of ATPase activities of full-length (green) and truncated (black) Wzm–Wzt. Shown are apparent catalytic rates in detergent-solubilized and liposome-reconstituted states. Data points represent the mean of three independent repeats with s.d. CPS, counts per second.


a
TM1
I25
IF
GH
90°
TM5
S44
I68
A59
TM2
T188
I191
TM1
I178
b
closed
open
c
kDa
WT
S44C/
A59C
A59C/
I25C
I68C/
I178C
T188C/
I191C
50
37
25
20
D
M
ADP/Mg + – + – + – – + – – + + –
Co-Phen + + + + + + – + + 2. – – –
STT – – – – – – – – – – – + +
NEM 2. 2. 2. 2. 2. 2. 1. 2. 2. 1. 1. 2. 2.

Extended Data Figure 9 | Model of the Wzm–WztN closed conformation. a, Rigid body alignment of the Wzm–WztN transporter halves with the corresponding NBDs of the closed WztN dimer structure. The closed WztN dimer is shown in grey, and Wzm–WztN is coloured in red and green for Wzm and cyan and blue for WztN. Residues replaced with Cys are shown with spheres at their Cα carbons. Observed disulfide cross-links are indicated with a dashed line. **b**, Cartoon illustration of the open to closed transition of the transporter. **c**, Disulfide cross-linking of Wzm protomers. Purified Wzm–WztN transporters harbouring the indicated Cys mutations were oxidized with either copper phenanthroline (Co-Phen) or sodium tetrathionate (STT), blocked with *N*-ethylmaleimide (NEM), and analysed by western blotting against the N-terminal Wzm Flag-tag. Experiments were repeated three times with similar results. M and D, Wzm monomer and dimer.

Extended Data Table 1 | Crystallographic data collection and refinement statistics

	WzmWztN	WzmWztN (Hg)	WzmWztN-T128C (Hg)	WzmWztN (Se-Met)	WztNBD (monomer)	WztNBD (Hg)	WztNBD (dimer)
Data collection							
Space group	$P4_132$	$P4_132$	$P4_132$	$P4_132$	$P3_121$	$P3_121$	$P3_121$
Wavelength (Å)	0.9895	1.0052	1.0052	0.9895	0.9895	1.0052	0.9895
Cell dimensions							
a, *b*, *c* (Å)	228.1, 228.1, 228.1	233.5, 233.5, 233.5	232.0, 232.0, 232.0	230.8, 230.8, 230.8	96.2, 96.2, 60.9	97.0, 97.0, 60.9	97.8, 97.8, 104.2
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 120	90, 90, 120	90, 90, 120
Resolution (Å)	24.9–3.85 (4.22–3.85)*	30.5-8.50 (9.51-8.5)	29.46–7.09 (7.93–7.09)	24.89-5.21 (5.82-5.21)	31.5–2.05 (2.11–2.05)	42.0-3.69 (4.04-3.69)	39.25–3.51 (3.84–3.51)
R_{merge}	0.23 (2.19)	0.12 (0.18)	0.12 (1.16)	0.19 (1.38)	0.08 (1.22)	0.09 (0.15)	0.16 (0.72)
R_{pim}	0.08 (0.73)	0.01 (0.02)	0.03 (0.26)	0.04 (0.31)	0.04 (0.61)	0.02 (0.03)	0.08 (0.35)
$CC_{1/2}$†	0.991 (0.43)	0.999 (0.99)	0.997 (0.81)	0.997 (0.78)	0.997 (0.82)	0.998 (0.99)	0.993 (0.66)
$I / \sigma I$	7.6 (1.3)	45.0 (33.5)	20.4 (3.0)	12.8 (3.0)	9.5 (1.5)	30.3 (20.8)	8.0 (2.8)
Completeness (%)	99.6 (100.0)	97.8 (100.0)	97.5 (96.1)	99.1 (100.0)	98.3 (99.3)	99.9 (100.0)	99.9 (99.9)
Redundancy	9.6 (9.9)	76.7 (81.6)	20.3 (19.5)	19.0 (20.1)	4.8 (4.7)	20.9 (20.9)	5.3 (5.4)
Refinement							
Resolution (Å)	24.9-3.85				27.8-2.05		27.1-3.5
No. reflections							
Total	36222				37942		13640
R_{free}	1824				1766		722
R_{work} / R_{free} (%)	25.8/32.1				20.2/23.4		24.5/30.5
No. atoms							
Protein	7962				1919		3579
PEG-400					204		
B-factors (Å^2)							
Chain A	172.9				67.3		107.8
Chain B	184.5						128.2
Chain C	130.3						
Chain D	145.8						
PEG-400					68.4		
R.m.s deviations							
Bond lengths (Å)	0.005				0.008		0.005
Bond angles (°)	0.856				0.859		0.813

*Values in parentheses refer to the highest-resolution shell.

†Correlation between intensities from random half-data sets[40].

POSTDOCS Better pay and benefits still needed, says advocacy group **p.369**

BLOG Personal stories and careers counsel **http://blogs.nature.com/naturejobs**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

ILLUSTRATIONS BY SALLY ELFORD/GETTY

HUMAN BEHAVIOUR

A kinder kind of science

Many researchers are calling for an end to the dominant winner-takes-all approach.

Scientists in New Zealand held the first 'Kindness in Science' workshop in December 2017 at the University of Auckland, hoping to kick-start a movement that will offer a kinder, gentler and more inclusive scientific culture. The group's mantra is "Everyone here is smart and kind — don't distinguish yourself by being otherwise."

At a time of great global divisiveness, moves are afoot to make the research culture more welcoming, respectful and responsible. Kindness, the workshop participants argue, should apply to the tone of peer review, to conference behaviour and to laboratory etiquette, among other areas. The winner-takes-all model is not the only way to make big breakthroughs in research, they suggest.

The group's working definition of kindness in science is "an inclusive approach that fosters diversity, respect, well-being and openness, leading to better science outcomes".

Nature interviewed seven researchers at various career stages to ask what such a culture shift might mean for them.

TAMMY STEEVES

Harness the power of the collective

A leader of the New Zealand Kindness in Science movement and a conservation geneticist at the University of Canterbury in Christchurch.

The idea for the movement came when three of us were reflecting on an essay about kindness in science, written in September 2016 by Emily Bernhardt, who at the time was president of the Society for Freshwater Science. Responding in part to sexist comments made to female junior researchers in her field, she urged colleagues to "rack up acts of intentional scientific kindness".

For me, as a mentor, this is about making space at the science table, creating an inclusive place for early-career scientists. Why does it matter? I believe that the inclusion of diverse perspectives is critical because it brings fresh approaches to tough scientific problems.

About a month after I read the essay, I noticed an opinion piece in *Nature* (T. Serio, *Nature* **532,** 415; 2016) on subtly sexist remarks towards women in science. The feedback it generated — towards a scientist who was talking about her own experience — was vile. Many commenters dismissed her examples of microaggression as misunderstandings. One accused her of being an "oversensitive damsel". I thought, "We have something positive to contribute here."

My colleagues and I want to get a global movement going by starting local. We envisage a diverse collective of scientists leading a culture shift that embeds kindness in how scientists work and how science is conducted. I see it as a shift away from empire building towards village growing. Currently, the vast majority of the science pie rewards the building of empires — that is, the model that has scientists clambering over one another to reach the top. That model can lead to amazing science, but it ▶

leaves only a sliver for people who approach and do their science in a very different way.

My intent is not to convince empire builders not to build empires, but rather to show the science community that there is another way. And I'm not talking about growing villages where everyone's singing 'Kumbaya' and holding hands. Rather, I mean harnessing the power of the collective to achieve better science outcomes. Collective efforts are rarely rewarded. That needs to change.

For my own research group, it works for us to run it as a collective. Our meetings allow students and postdocs to say what they want to achieve, to be in control of what they do and to take ownership of their research. Also, when things aren't going to plan, the collective response isn't, "You're doing it wrong", but "How can we help?" We have really high standards, and we meet these as a group by helping one another to meet them individually. If I want to foster a community of smart and kind scientists, I've got to give them the space to be just that.

I believe that kindness in science will lead to better science outcomes. With more scientists working in a truly inclusive way, we will achieve more.

JAMES ATARIA
Collaboration is crucial

Ecotoxicologist at the Bio-Protection Research Centre, Lincoln University, Christchurch, New Zealand; deputy director, Ngā Pae o te Māramatanga, New Zealand's Māori Centre of Research Excellence in Auckland.

Science is a cut-throat enterprise. But I've always been of the view that we get so much more done working together than against each other. From a cultural perspective, as a Maori researcher, I'm all about collaboration and working together. So I really see the notion of kindness in science as being a positive thing for bringing more emerging Maori researchers into science.

There's room in science for being more collaborative. In New Zealand, our government-funded Crown Research Institutes are expected to turn a profit. In a system where financial viability is all-important, science can take a backseat in some instances. Such a system also creates an environment in which competition runs rife.

But I see benefits in a more collaborative environment, with better outcomes across the board. For example, working on a Mataura River project in a southern region of New Zealand, at the outset we decided to involve local Maori organizations, as well as representatives of regional and central government, to address their concerns about the river ecosystem.

In the past, researchers might have just put their heads down, done their research and published it or put out a report. But our research was much more embedded in the community.

We decided to sleep, eat and work in the headquarters of the local *rūnanga*, or Maori governance organization. We had opportunities to engage with parts of the community that we wouldn't normally have any reason to associate with, and they asked us to explain what we were doing. When researchers have a really strong social or community connection, they can see why they are doing the research.

Kindness is quite an evocative term, but I see it come through when researchers experience how their work is changing a community. Likewise, from the community's perspective, being at the decision-making table and co-generating research is empowering, and is a form of kindness. We've got these concerns, you've got expertise: how can we pair them together? Collaboration with communities can both create conditions for kind science and produce good scientific outcomes.

JAMES DOTY
Sympathy lets creativity flourish

Neurosurgeon, and founder and director of the Center for Compassion and Altruism Research and Education at Stanford University School of Medicine in Palo Alto, California.

There's an increasing body of evidence that how we behave towards others matters. Unfortunately, what is required to succeed in academic science is antithetical to being kind.

A lot of people will refer to Darwin, saying it's dog-eat-dog and that only the strongest survive. But what Darwin really said is, it's the survival of the most sympathetic (go.nature.com/2lsloz6). In every society, it's the efforts of those who are most sympathetic and caring that result in the long-term survival of the species.

When you care, or demonstrate caring behaviours, the hormone oxytocin is released and gives the brain a sense of calmness and connection. When you feel threatened, the fight-or-freeze system is activated instead, and the executive control function and creativity areas of your brain shut down. As a result, you don't often pick the best behaviour — you pick one that you think will allow you to survive that moment.

In the hypercompetitive environments of academia and business, every step involves ruthlessness. The competitiveness never ends. But people's stress response is always engaged. This has a huge negative effect on their health.

We know that when the work environment of a business becomes kind, compassionate and thoughtful, it reduces stress in employees and results in people being more open, more discerning and less judgemental. The same should apply in academia, enabling more creative, thoughtful research and making researchers more productive.

DAVID COLQUHOUN
'Publish or perish' is a foolish fetish

Biophysicist at University College London, UK.

Excessive competition between individuals, journals and universities has reached levels where it's endangering the reputation of science and hurting people. Several years ago, I heard a BBC morning news programme in which the host asked a researcher, "Is this a real breakthrough or are you applying for a new grant, or are you starting a spin-out company?" That's a terrible reputation for science to have.

Pressure to publish, whether there's anything to say or not, is an incentive to cut corners, and occasionally to be outright dishonest. It is common to have big labs headed by people who can only be described as jerks. Graduate students and postdocs are often used as slaves, working for the glorification of their lab head and their university. It's not unknown for junior people to be bullied by lab heads to get a particular result. The 'publish or perish' obsession reduces the quality of published work, and can even lead to suicides. And the real tragedy is that it's based on metrics that are nonsense. Citations don't measure the quality of research, and rankings don't measure the quality of universities.

The main problem is that there are too many people doing science now. There is bound to be bad behaviour when the available funding doesn't match the number of people who want it. Funding isn't likely to increase, so what can be done? We need to reduce the number of institutions doing research. That would free up money for the later stages of research and lower the numbers of people applying for grants. However, it's not a very politically viable option.

But we do need a change in culture. There are two reasons to stop the incentives that encourage the jerks: it would increase the quality of research, and it would contribute to the sum of human happiness.

EMILY BERNHARDT

Tone down the criticisms

Ecologist at Duke University in Durham, North Carolina, and author of an essay on kindness in science.

When I began drafting my essay (go.nature.com/2czt3pc), I intended to write a scathing one. But then I realized we needed something more positive. We all need to be more attentive to being kind. The low-level racism and sexism that exist in science do real harm.

Senior people should be calling out bad behaviour: "I think that's a little over the top or unkind." That will make people step back. And saying "Can you back up that statement?" can be quite effective among peers.

Unkindness is rife in the review process, and journal editors can do a lot to help by asking reviewers to tone down their criticisms. I've seen students destroyed by a mean review that insults their intelligence or writing, rather than focusing on the science. First papers are such important things for young scientists, and that first review feels like a statement on their abilities as a human being and a scholar.

There's this idea that it's OK to be an awful person as long as you are brilliant. But there are tons of people who are generous with their time or positive energy and who make academia work better.

BINYAM MOGESSIE

Make mentoring matter

Cell biologist, University of Bristol, UK.

As a new principal investigator, the most important thing for me is to be a member of the lab and not 'the boss'. A lab should be a place for the growth and development of everyone who joins it. If someone needs my support, not just for trouble-shooting experiments but because science is very challenging and demotivating at times, I will tell them, "You should take a week away, go to a conference or give a seminar, and get excited about the science again." As a principal investigator, you need to acknowledge that you have a responsibility for every person you hire.

When you move from a PhD to a postdoc or academic position, no matter how hard you work, you still need a lot of mentoring. The person you are working for must think about your career progression and the things you should be doing — even if that means just taking 10 minutes to sit down with you and find out what you are interested in doing.

Nothing is definite in this business. You can't have an edge on people competing for the same job if you do not know what to expect at the next level. That's when an adviser or a mentor has to step up. If someone is willing to share that information with you, it is really kind.

STEPHANIE GALLA

Build bridges, don't burn them

PhD student in biology, University of Canterbury, Christchurch, New Zealand, and a founder of the Kindness in Science movement.

Being an early-career scientist is a cool time. It's when you get to explore what kind of science you want to study and what kind of scientist you want to be. It sets up the trajectory of your career.

But some things make me ask, do I really fit in here? There are long-lived lab rivalries that affect the quality of the science. That's disheartening. I've also met people who are more possessive about their science and not willing to share their research wisdom, data or code.

Overall, I've been fortunate to work with very kind scientists. I've just come from a meeting with government agriculture researchers who invited me to their lab group to talk about bioinformatics, and they were willing to share their hard-earned wisdom with me.

This helped me to make leaps and bounds in my own research and also led to mutually beneficial conversations on how to best approach shared research questions.

I think kindness is the path forward. I don't want to be a bridge burner, but a bridge builder. That's going to lead to better science.

INTERVIEWS BY KENDALL POWELL

Interviews were edited for clarity and length.

POSTDOCS

Support slowly grows

Academic institutions in the United States have helped to improve life for postdoctoral researchers but changes are still needed, according to a 3 January report from the National Postdoctoral Association (NPA) in Rockville, Maryland, which represents postdocs in the United States and Canada.

Supporting the Needs of Postdocs recommends that postdocs receive higher compensation, equal benefits regardless of how a researcher is classified or funded, and more-generous parental leave.

The report collated results from a 2016 survey completed by 102 of the 190 institutional NPA members that maintain a postdoctoral office on campus. The survey results, published in partnership with Sigma Xi, a researcher association in Research Triangle Park, North Carolina, indicate that 94% of member institutions require that new postdocs and other recruits learn about appointment policies and resources, and that 85% of institutions have an orientation programme that outlines services and amenities available to postdocs.

Postdoc pay rates, however, are less consistent across member institutions, despite federal legislation passed in 2016 that compels employers to either raise the minimum salary for all US hourly workers to US$47,476 a year or offer overtime pay. Survey responses indicate that 77% of institutions pay that rate or are raising their minimum compensation to that level. Just 36% of institutions require annual stipend increases, 43% recommend it and 21% have no policy on the matter, the report says.

Most postdocs receive health-insurance benefits and paid time off, but postdocs who have their own funding often lose access to institutional benefits. This is a continuing point of contention, and the NPA urges institutions to address it.

The report recommends that institutions determine postdoc needs more effectively by gathering information on diversity, disability and disadvantaged backgrounds. It also calls for universities to maintain contact with postdocs after they leave, so as to develop a comprehensive alumni network and to track career pathways. Currently, 45% of institutions carry out exit surveys, and 28% track their postdocs after they leave the institution.

Since 2000, various societies and organizations have published reports on the importance of postdoctoral researchers to the US scientific enterprise and how postdoctoral training can be improved.

CHOCOLATE CHICKEN CHEESECAKE

A taste of success.

BY M. J. PETTIT

Bryant bit his lip as three-Michelin-starred chef Jean Christophe assessed the evening's final plate. The pinkish meat, coarsely butchered, sat in a pool of steaming liquid. Chef bent over and wafted the aroma, catching himself before he recoiled. "Tell us about your dish, Nostradamus."

The avatar's hydraulics hissed as it craned its head into focus, giving the cameras an almost human smile. A rainbow of stains and scorch marks littered its chef's jacket. "I've prepared for you a carnival of chicken and pork sashimi." A pair of icy blue lamp-lights pulsed as Nostradamus listed the mountain herbs strewn about the plate.

"Bathed in raspberry coulis?"

"I'd call it more of a drizzle, chef."

Chef gave a blackened smear a tentative poke with his fork. "Is this *mole* sauce?"

"No. My sundrenched mayonnaise got a touch over-ripe but I did have time to plate it this week."

A forkful suspended before his mouth, Chef Christophe shot a pleading look off camera.

"You don't like my dish," the ever-prescient Nostradamus observed.

"No," Chef fumbled for words, "just taking a moment to savour your … creativity."

"But you haven't tasted it yet."

Bryant squeezed Rita's delicate hand. His beloved for the past three episodes had stumbled about the kitchen ever since her amuse bouche mysteriously turned rancid. It was down to her or Nostradamus. The real mystery was the machine's continued survival given its monstrous palate and brutal technique.

Then Chef Christophe coughed out those impossible words: "Genius, utter genius."

Back in the dorms after kissing Rita goodbye, Bryant decided he needed to confront someone, anyone. He understood now. He was never meant to win. But he wouldn't go quietly. He couldn't. He still had a chance to return for the reunion special as fan favourite.

He found the evening's other runner-up alone, the remaining occupant of Team Virgo's quarters. Sally dangled her legs from the upper bunk as she unwound with a Bud Lite Lime. With her cropped white hair and sinewy frame, she looked like a seasoned warrior. Rumour was she'd once been military, before joining the cast back in season one.

"Too bad about Rita," Sally said. "I'll miss her."

ILLUSTRATION BY JACEY

Bryant waved away the phony regret. "I get it now," he said, "why it's so much harder in person than it looks on TV."

"Tell me about it. Nostradamus is one tough cookie."

"But how does it compete without a sense of taste or smell?"

Sally shrugged before taking another slug of her beer. "We're playing against something akin to the Internet. It has to win. You get that, right? Besides, Nostradamus made a great comeback tonight. Such a brave flavour profile."

"They had to pump the guest judge's stomach."

Bryant faced the not-so-hidden camera streaming their conversation for the after-show fans.

"That could have been anything."

"Rita didn't deserve to get sent home."

"Sorry they eliminated your girlfriend, but she shouldn't have used frozen scallops. Destroys their texture."

"It was a deep-freezer challenge."

"Whatever, man-bun."

Bryant winced. How had that awful nickname stuck? "I've figured it out, why each season starts again with eleven new contestants but keeps the two of you."

Sally shrugged. "We have a history. Nostradamus sees me as its biggest threat. Gets a thrill each time it defeats me."

"That's not it. I saw you sabotaging Rita tonight."

Sally eyed the camera. "Listen kid, Nostradamus held on fair and square."

"You poured boiling water into her ice cream maker. Swapped salt for her sugar while the host distracted her with an interview."

"Not my fault she let the pressure get to her."

"To think I used to root for you," Bryant said. "How much do they pay you to make that machine appear more than human?"

"You don't know what you're talking about."

"Why can't you admit this whole thing is fake?"

"Fine, but what were you expecting? We're on TV."

"Is it true, Sally?" Nostradamus asked. Neither of them had noticed it eavesdropping at the door. "I don't win fair and square."

The avatar's head drooped.

Sally leapt from her bunk to console it, but Nostradamus bolted away.

"Shit," Sally turned on Bryant. "You couldn't drop it, could you?"

"No, I came into this competition looking to win."

"Really? After 18 seasons, you thought I'm going to defeat Nostradamus. It's the most powerful consciousness on Earth."

"And a terrible cook."

"With an incredibly fragile ego," Sally said. "Think I want to keep competing? We all had our part to play. Even Rita understood that."

"So why mess with her?"

"I couldn't trust the judges. They're only human," she said as she paced the room. "Taking care of Nostradamus was my responsibility."

"I don't get it. You're talented. You could win this thing if you tried. Put an end to it all."

Bryant started in the direction of an unexpected rumble of thunder.

Sally gulped the last of her beer. "I remember when life was simple, when Nostradamus was just a neural net scraping Pinterest boards and food blogs. The first iteration spewed out nonsense recipes like 'chocolate chicken cheesecake.'"

Bryant watched as the lights flickered.

"It stored the memory somewhere because after self-awareness, Nostradamus declared its intentions to eliminate us," Sally continued. "Thankfully, it meant on a damned cooking show. My superiors could hardly say no, given the circumstances. Kept it distracted from the doomsday scenario. Well, at least until you came along, man-bun."

The dormitory shook with a second thunderclap. Bryant ducked under the bunk as fissures ran down the concrete walls.

Sally sounded resigned as chunks of plaster fell around her. "So how are you feeling about our next judges' table?" ■

M. J. Pettit *is an academic and writer who divides his time between Toronto, Canada, and Manchester, UK.*

naturejobs

CAREER GUIDE CHINA

NATURE, VOL. 553, NO. 7688 (18 JANUARY, 2018)

COVER IMAGE: CASPER SHAW IMAGE/GETTY

Those new to China are often overwhelmed by its size and scope. More than 100 Chinese cities now have more than 1 million residents, and 55% of the nation's 1.38 billion people live in urban areas. When the capital Beijing became overcrowded, the government began building a new city, twice the size of Manhattan, next door.

The transport is equally supersized. If you joined China's railway lines together, they would loop around Earth twice. Shanghai's metro system is the longest in the world, and Beijing has the planet's busiest subway, transporting more than 10 million passengers each day. The world's biggest airport is being built 68 kilometres south of Beijing — it will eventually have seven runways.

To scientists contemplating a move to China, the nation's scientific ambitions can seem similarly daunting. By 2049, the centenary of the founding of the People's Republic of China, the country plans to be a world-leading science and technology power. To this end, China "needs the strategic support of science and technology more urgently than ever before", said President Xi Jinping in 2016.

Research into everything from space exploration, quantum communication and brain research to big data applications, clean energy and robots is part of China's plan to evolve from a low-cost manufacturing hub into a modern, innovation-driven economy. As one Chinese neuroscientist told us (see page S4): "Scientists must link their research plans to the government's national demands, rather than purely their own academic interests." Spending on research and development in China is already far above countries with similar gross domestic product per head. It remains behind only the United States in terms of total spending, a situation that is likely to be reversed by 2020.

This is a country where high levels of investment and ambition are creating exciting career opportunities for scientists. Recruiters are looking for researchers from all disciplines to fulfil China's dreams to become a world leader in artificial intelligence (S10), biotechnology (S19) and commercial innovation (S28).

We also look at the realities of moving a career to China (S14). We talk to scientists who have returned home after living abroad, or made the move with little experience. Their experiences are diverse, but one point is clear: China's science landscape is transforming rapidly. Those interested in being a part of it should start their search now.

Sarah O'Meara
Guest editor

CONTENTS

S2 WHY CHINA NEEDS YOU
Scientific research is key to the nation's future

S4 HOW TO FIND A JOB
Generous funding has attracted researchers to China, but what are the hurdles they face?

S8 WHAT IS CHINA'S THOUSAND TALENTS PLAN?
A guide to the famous recruitment scheme

S10 CHINA'S AI DREAMS
World leader in artificial intelligence by 2030?

S14 MOVING TO CHINA
The triumphs and difficulties of relocation

S19 BRACED FOR THE BIOTECH BOOM
Government investment is building a life-sciences powerhouse

S24 PUTTING SCIENCE ON THE MAP
Large-scale science projects across China

S28 THE INNOVATION HUBS THAT DRIVE CHINA
China's manufacturing sector is evolving fast

S32 DRAWN TO SHANGHAI'S SCIENCE SCENE
Cosimo Bambi on his move to the megacity

CHIEF CAREERS EDITOR: David Payne. **EDITORIAL**: Sarah O'Meara, Jack Leeming, Lewis Packwood. **ART & DESIGN**: Mohamed Ashour, Kate Duncan, Madeline Hutchinson. **PRODUCTION**: Kay Lewis, Ian Pope, Dipti Shah. **MARKETING**: Claire Jones, Seyi Ibidunni. **SALES**: Nicole Yu, Joan Yin, Janet Cen. **DIRECTOR, GLOBAL CAREER SERVICES**: Nils Moeller. **PROJECT MANAGEMENT**: Rebecca Jones, Sharon Wang. **EDITORIAL DIRECTOR**: Stephen Pincock. **ART DIRECTOR**: Wojtek Urbanek. **PUBLISHER**: Richard Hughes. **EDITOR-IN-CHIEF**: Philip Campbell.

NATUREJOBS CAREER GUIDE
The Naturejobs Career Guide China 2018,
a supplement to Nature, is produced by Nature
Research, the flagship science portfolio of Springer
Nature. Naturejobs (naturejobs.com) is the global jobs
board and career resource for scientists.

NATURE EDITORIAL OFFICES
The Campus,
4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7833 4000
Fax: +44 (0)20 7843 4596/7

CUSTOMER SERVICES
To advertise with Naturejobs, please visit
naturejobs.com/support or email naturejobs@nature.com
feedback@nature.com

WHY CHINA NEEDS YOU

Scientific research is at the heart of the country's plan to transform its future.

BY DENISE HRUBY

When Raymond C. Stevens moved to Shanghai in 2011 for a nine-month sabbatical, he was curious. After spending his career in the United States, establishing a laboratory at the University of Southern California in Los Angeles and three biotechnology start-ups, the chemist sensed that China might be the next research frontier.

"I wanted to get a refreshing perspective," says Stevens, now the director of the iHuman Institute at ShanghaiTech University and founder of another, China-based start-up. "I was curious about science in China, what drug discovery was like, and wanted my three children to experience life in China since I felt it might open opportunities for them later in life." Although Stevens enjoyed life in China, he and his family wanted to be closer to their relatives in the United States, and, six years later, decided to move back. Stevens continues to spend almost half his year working in Shanghai.

Whether you're a tenured professor or just starting out as a postdoctoral researcher, it's hard to miss China's vast science and technology ambitions. During the opening of the 19th National Congress of the Chinese Communist Party last October, President Xi Jinping affirmed the need for China to become "a nation of innovators". In 2019, China is predicted to surpass the United States as the world's largest investor in research and development (R&D).

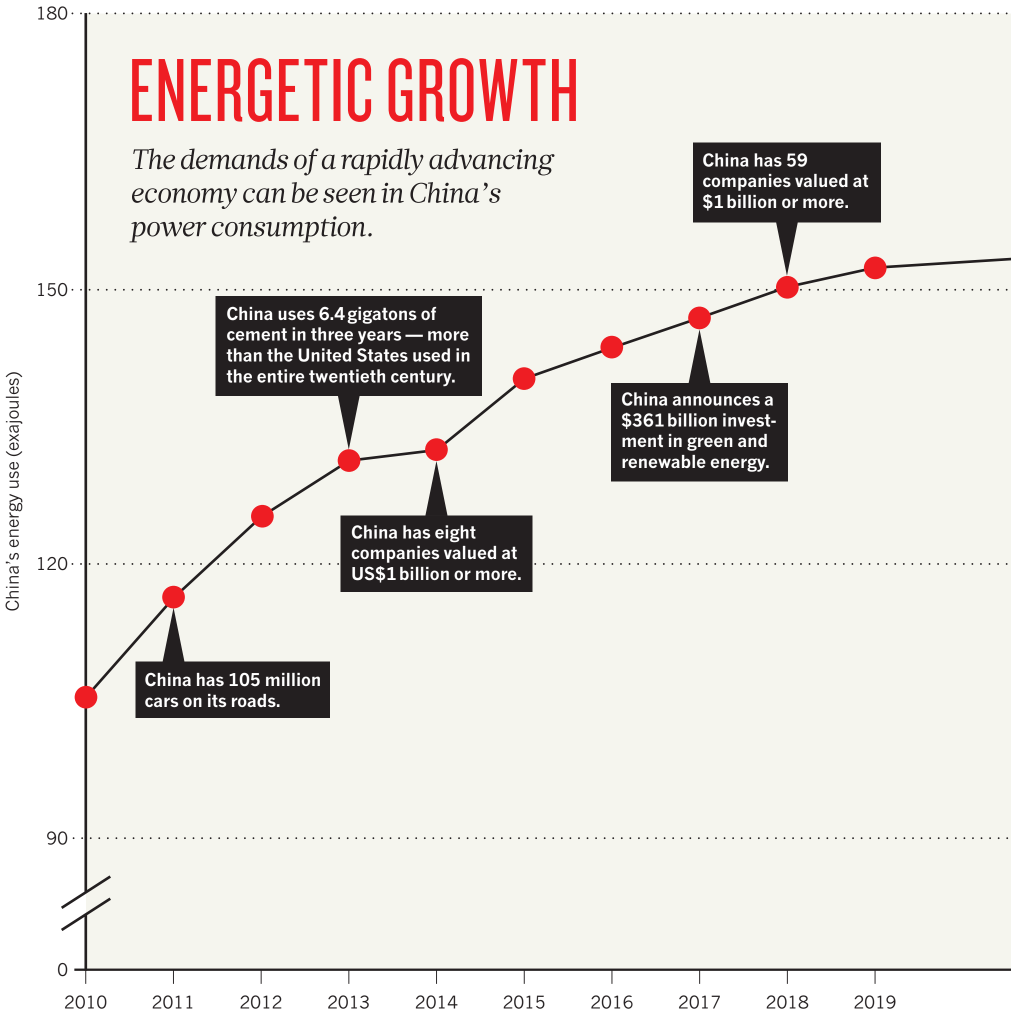
Over the past decade, China's leaders have focused on transforming a nation of farmers and factory workers into a highly skilled workforce. In 2006, China announced a 15-year science and technology plan, which set targets for everything from spending levels to the number of patents that must be filed. The aim was to transform China from a low-cost manufacturing hub into an economy driven by domestic innovation. In 2016, China's leaders went further, stating that by 2050 the country will be a world-leading scientific powerhouse (see 'Energetic growth').

"Science is culturally well appreciated in China, perhaps comparable to sports in the United States," says Stevens. "Students are filled with incredibly strong desire and sense of curiosity about science. For me this is the most exciting reason to be working in China today."

China's investment in R&D is crucial to the country's stability. As the country's economy matures, its growth rate is slowing. The economy is expected to grow no more than 6.4% in 2018 — a far cry from the rapid gains of 10% per year in recent decades. Coupled with this economic deceleration are the twin pressures of growing labour costs and a shrinking workforce. Owing to the country's historically strict family-planning policies, by 2050, more than one-quarter of China's population will be over 65, and only around one-half will be of working age.

INVESTMENT FOR THE FUTURE

Well aware of the need for change, the country's leaders are putting money behind new businesses. Start-up investment funds run by local authorities across the country totalled US$332.6 billion in 2015, and the plethora of business incubators across the country helped to produce 223,000 companies by the end of 2016. The consultancy firm McKinsey says that innovation could contribute as much as 50% of the country's gross domestic product (GDP) growth by 2025. By 2020, China wants to have more than 10,000 domestic incubators, accelerators and other start-up havens, and 100 abroad. From designing new semiconductors and chemical products to re-imagining the role of artificial intelligence in everyday life, China will be relying on its growing pool of talented scientists and entrepreneurs to come



ENERGETIC GROWTH

The demands of a rapidly advancing economy can be seen in China's power consumption.

China has 105 million cars on its roads.

China uses 6.4 gigatons of cement in three years — more than the United States used in the entire twentieth century.

China has eight companies valued at US$1 billion or more.

China announces a $361 billion investment in green and renewable energy.

China has 59 companies valued at $1 billion or more.

SOURCE: US ENERGY INFORMATION ADMINISTRATION up with ideas, bring them to life and take them to market.

State-of-the-art research facilities are being built across the country, and scientists say research funding is prolific. “I feel that in North America, the pace is very slow. In China, there is all the support and all the funding you need, so the pace is very fast,” says Zhongqi Shao, who moved to China to work as vice president of vaccine research at the vaccine developer CanSino Biologics in Tianjin, after having worked for pharmaceutical companies in Canada for decades.

The strong support for science and technology is also a means for China to address pressing social problems; 43 million people are living below the country’s poverty line, according to government figures. Farmers make vital income by selling everything from trinkets to furniture and dried fruit on China’s giant e-commerce website Taobao. On the streets of the country’s cities, young entrepreneurs have transformed the commute of millions of Chinese with the introduction of bike-sharing apps. They hope to reduce air pollution as well as traffic jams.

Technology is also tackling problems arising from China’s rapidly ageing society. Companies are counting on artificial intelligence to replace millions of unfilled low-skilled and labour-intensive jobs. Like everyone else, China’s elderly are embracing technology: robots are not only being deployed to assist in medical care, they’re also seen by residents in care homes as a means to ease their loneliness. They encourage the little machines to dance and play karaoke songs.

PLANNING AHEAD

China has been run by the Communist Party for almost 70 years, and the resulting stability has enabled politicians to set and execute long-term goals. The nation’s science and technology priorities are plain to see in the latest five-year development plan — the thirteenth since 1953. The government’s priority is to foster innovation by increasing the share of GDP spent on R&D to 2.5% by 2020, up from 2.1% in 2015, which would bring it on par with the average spend by countries in East Asia and the Pacific region. If China reaches this target, it will spend an estimated $1.2 trillion on R&D over the course of the plan.

China is already the world’s second-largest investor in technological innovation, after the United States, and spent more than $200 billion on R&D in 2015. Since the turn of the century, China has increased tertiary education spending and created state-of-the-art research facilities — investments in human capital seen as vital to drive innovation. Under the new plan, policymakers hope to increase the number of staff directly working in R&D — from bench engineers and scientists to their managers — to 60 per 10,000 employees, up from 48.5 in 2015. The plan also suggests that 10% of China’s population should have a scientific degree by 2020, up from 6.2% in 2015. But employers say that Chinese graduates still lack the soft skills necessary to generate new ideas, such as analytical thinking and an ability to communicate effectively.

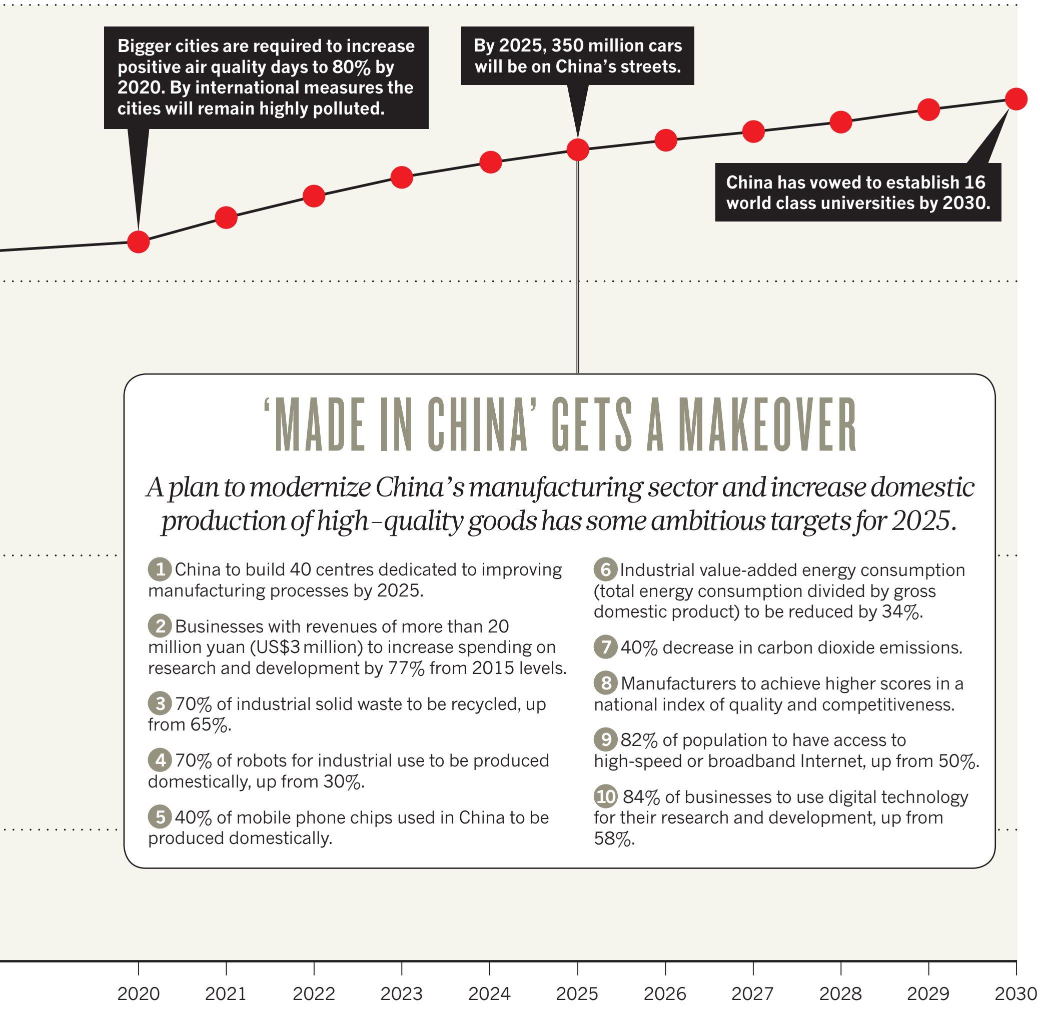
> **“THE INVESTMENT IN HUMAN CAPITAL HAS BEEN SIGNIFICANT.”**

Slowly, however, past investments are paying off, says Cong Cao, a professor of Chinese studies at the University of Nottingham Ningbo in China. “The investment in human capital has been significant, with Chinese universities having turned out more and increasingly high-quality graduates,” Cao says.

To bridge the gap between research and commercial use, the government is investing heavily, and has created a national small- and medium-sized enterprises development fund, with $9.4 billion slated for the development of private start-ups. China has also invested in several online services in the travel and tourism industry. Innovation parks and centres are sprouting up even in smaller cities; some offer free rent to individuals for up to five years. University graduates are being offered financial incentives by city governments, such as free rent and tax cuts, to take a leap of faith and start a business.

Although China still lags behind high-income economies when it comes to certain areas of innovation, such as designing car engines or creating new drugs, the policies seem to have had tangible results. According to the report *The Global Innovation Index 2017*, which measures dozens of indicators ranging from patent filings to education spending, China leapfrogged Estonia and Australia to go from 25th to 22nd in 2017.

Perhaps more tellingly, state media reported that in 2015, double the number of students intended to start their own businesses compared with the previous year. Whatever they start will have to be flexible — China is changing fast. ■



‘MADE IN CHINA’ GETS A MAKEOVER

A plan to modernize China’s manufacturing sector and increase domestic production of high-quality goods has some ambitious targets for 2025.

1 China to build 40 centres dedicated to improving manufacturing processes by 2025.

2 Businesses with revenues of more than 20 million yuan (US$3 million) to increase spending on research and development by 77% from 2015 levels.

3 70% of industrial solid waste to be recycled, up from 65%.

4 70% of robots for industrial use to be produced domestically, up from 30%.

5 40% of mobile phone chips used in China to be produced domestically.

6 Industrial value-added energy consumption (total energy consumption divided by gross domestic product) to be reduced by 34%.

7 40% decrease in carbon dioxide emissions.

8 Manufacturers to achieve higher scores in a national index of quality and competitiveness.

9 82% of population to have access to high-speed or broadband Internet, up from 50%.

10 84% of businesses to use digital technology for their research and development, up from 58%.

Denise Hruby *is a writer and editor based in Shanghai, China. Additional reporting by* **Shannon Ellis.**

HOW TO FIND A JOB

China's national science goals and high funding levels are attracting researchers, but there are barriers to overcome.

BY HEPENG JIA

Theoretical chemist Jeffrey Reimers had been working in Sydney, Australia, for 28 years when he decided to relocate. In previous years he might have moved to the United States, but he decided that China's deep pockets and ambitious research targets could better support his professional goals. "I wanted to do something that could not be done if I moved to the United States. I wanted the opportunity to expand my research, to ask some serious questions."

Fast-forward four years and Reimers is now the director of the International Center for Quantum and Molecular Structure at Shanghai University, a multidisciplinary research hub with start-up funding of 20 million yuan (US$3 million). Since opening it in 2014, Reimers and his cofounder, returnee physicist Wei Ren, have hired more foreign scientists than native Chinese researchers to join their team.

The recruitment drive reflects the country's desire to attract international expertise. "China has developed to a stage that it can and should attract excellent global science talents," says Xiao-Fan Wang, a professor of pharmacology and cancer biology at Duke University in Durham, North Carolina, who frequently advises China's leadership on science and technology policy. "Our leaders are aware that if China wants to compete for global leadership in the future, it is important to attract international talent as the United States did after the Second World War. They are aware that all important science papers are published in English and very few high achievers have made it without overseas experience."

Go online today and you'll find numerous job listings for research positions in China. Yet the country's hunt for overseas talent only started in earnest relatively recently. In the early 1990s, China's top governmental science bodies launched a succession of schemes designed to woo talent to its shores after decades of academic decline. China began recruiting skilled professionals from abroad as early as 1994. The earliest scheme, the Hundred Talents Program, attracted 2,145 scientists to China by the end of 2013. It was closely followed, in 1997, by the Changjiang Scholars Program to hire overseas Chinese academics, which brought 2,251 professors back by 2014. Both foreign and ethnically Chinese scientists were encouraged to apply, and the majority of successful candidates were junior faculty members.

CHINA AIMS HIGH

But the recruitment landscape has dramatically changed from welcoming nearly all foreign-trained PhD holders to primarily recruiting potential academic leaders or those with significant academic outputs. Hailiang Yu, a professor in the School of Mechanical and Electrical Engineering at Changsha-based Central South University, said that for the past three years his school has no longer been interested in recruiting foreign graduates — it only advertises widely for senior research staff.

In 2008, China re-focused its recruitment efforts on high-end academics and introduced the Thousand Talents Plan (see page S8), which offered incentives such as high salaries to Chinese-born, Western-trained professors who were willing to return to China. Two years later the scheme was extended to foreigners, and an additional Youth Talents Plan was created to attract promising researchers up to the age of 40. By mid-2017, the Thousand Talents Plan had attracted more than 7,000 scientists to China, including 2,900 under the Youth Talents Plan and 381 foreigners.

China's timing was prudent. Research funding in the West was badly hit by the effects of the global financial crisis in 2008, at the same time that China was pouring money into recruitment. The country promised overseas professors start-up packages of up to 20 million yuan, including annual salaries of up to 1 million yuan in addition to a relocation payment of 1 million yuan. Reimers was one of the programme's recruits.

Figures released by the Chinese government suggest that these programmes have jointly attracted nearly 15,000 scientists to the country — predominantly, but not exclusively, Chinese returnees. Former President Hu Jintao,

"ATTRACTING SENIOR TALENTS IS PLACED AT THE CENTRE OF CHINA'S POLICY AGENDA."

MIR156/ALAMY

More than 7,000 high-fliers have relocated to China under the Thousand Talents Plan.

who led China from 2003 to 2013, stressed the importance of cultivating talent to bolster the country's industrial development throughout his tenure. His commitment to building up China's pool of skilled researchers has been continued by current leader President Xi Jinping, who in 2017 said that in the next five years, 50 more collaborative projects would be set up between Chinese and foreign scientists to encourage "talent exchange". China will also invite 2,500 young foreign scientists on short-term research visits.

"Attracting senior talents is placed at the centre of China's policy agenda and hence becomes a central task," says Xiong Zhou, deputy director of human resources at Huazhong Agricultural University in Wuhan. Neuroscientist Lu Bai at Tsinghua University in Beijing adds: "Now, we can compete with many leading Western universities for smart young talents with comparable benefit packages."

Chinese-born scientist Weiwei Deng, a professor of mechanics at the Shenzhen-based Southern University of Science and Technology (SUSTech), says he was offered an annual 9.5 million yuan in start-up funding for three years. Deng was a tenured associate professor at Virginia Tech in Blacksburg before he joined SUSTech through the Thousand Youth Talents Plan in early 2017. "In these three years, I can do what I want without spending time on thinking about grant proposals," he says (see 'Who funds China's science?').

Researchers tend to find China's science spending much more lavish than what they're used to in the West. Zilong Qiu, a senior neuroscientist at the Shanghai-based Institute of Neuroscience of the Chinese Academy of Sciences, who returned to China from his postdoc position at the University of California, San Diego in 2009, says that many of his former colleagues can't get as much done as they'd like owing to a lack of money.

But scientists looking for funding in China must align their research with national priorities, he says. "China is still a developing country. To gain funding, scientists must link their research plans to the government's national demands, such as controlling diseases, rather than purely their own academic interests."

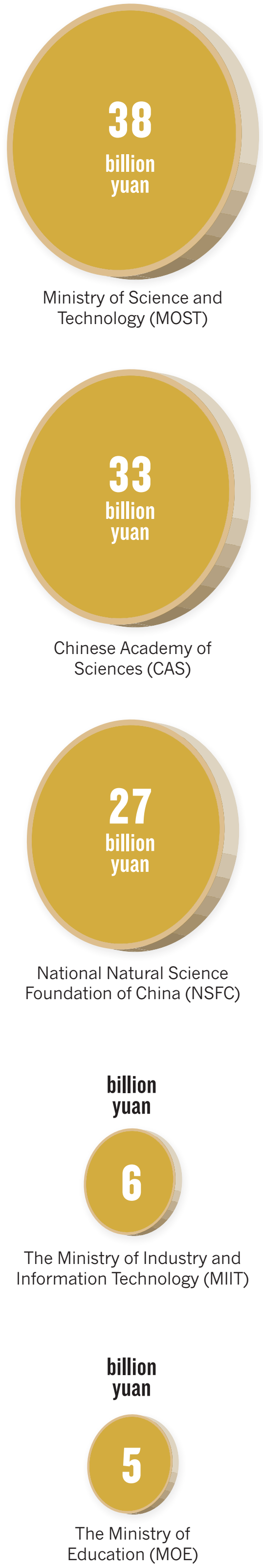
THE REALITIES OF JOB SEEKING

A common complaint among Chinese researchers is the blunt way that supervisors evaluate their performance. Institutions judge scientists on their publication record, often generously rewarding those whose work appears in high-impact journals. This pressure to publish has been associated with the spread of fraud among Chinese papers. For job seekers, the same evaluation method applies.

"The number of papers is the most ▶

WHO FUNDS CHINA'S SCIENCE?

China's funding organizations have deep pockets and are backed by the highest levels of government.



important criteria for Chinese universities to recruit Thousand Youth Talents candidates, but to many emerging research areas like mine, it is hard to publish a paper in a high-impact journal," says Guojun Shi, a cell biologist at the University of Michigan in Ann Arbor. Yet Lu is optimistic that the recruitment landscape is becoming more sophisticated. Leading universities, he says, are now keen to hire individuals who can be integrated into their research environment (see 'Where to look').

Another concern for job seekers is the part that nepotism plays in Chinese universities. Zhou notes that recruitment decisions are often reliant on personal connections. Scientists who have studied and worked abroad may struggle if they do not have the necessary patronage to secure funding and promotion. He recommends that Chinese institutions help newly recruited overseas scientists by arranging for them to work with senior scientists and in established labs.

There are no official guidelines for universities to follow when assimilating newcomers into the workplace, beyond the contractual obligations of the scheme. Reimers' lab, for example, hired an English-speaking administrator to help foreign staff. But this kind of support is not guaranteed.

FUTURE FOR FOREIGNERS?

Although China is committed to recruiting foreign talent (see 'In the fast lane'), the numbers of overseas scientists based in China remain small compared to the researcher population of 1.5 million.

According to Zhou, China's recruitment schemes struggle inside a culture that is not used to hiring foreigners. In 2015, just 240,000 foreigners were issued with work permits in China. By comparison, more than 530,000 temporary work permits were issued in the United States in 2016.

Language and cultural barriers, recruitment habits and huge supplies of Chinese nationals trained overseas are all factors explaining the few foreign scientists working in China, says Zhou. "It is only in recent years that Chinese universities began to advertise job openings in English in major international journals and academic websites," he says.

Although China offers high salaries to scientists recruited through its talent schemes, its average faculty salary is almost half the US equivalent. Ren, the codirector of Reimers' lab, says he regularly argues with his university's administration for higher salaries for foreign faculty members. The Chinese government is aware that low salaries do not encourage and retain talent. In November 2016, China's State Council issued guidelines for universities to increase the income of scientists through pay rises, greater rewards for research sold to companies and the use of grant funds to supplement salary.

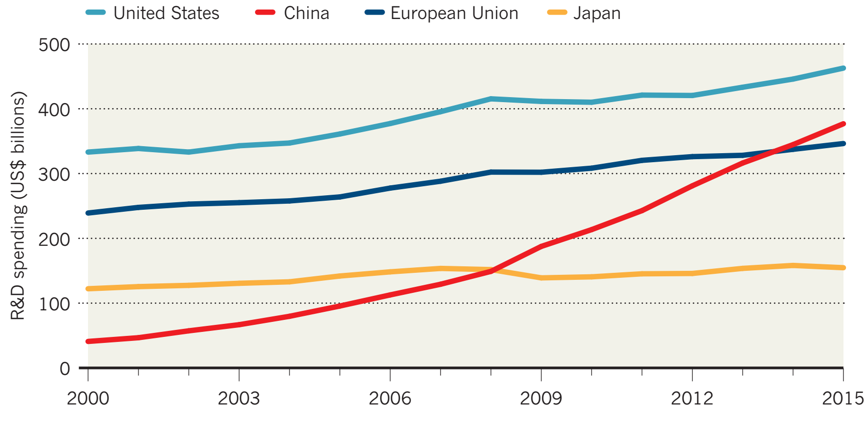
But there are still national policies to be overcome when it comes to recruiting foreign scientists. "Many of our policies and practices need to be overhauled if we want to massively attract foreign science and technology talents," says Huiyao Wang, director of the Beijing-based Center for China and Globalization, a think tank that advises the central government on talent policies. He notes that China issued 1,576 green cards, or permanent residencies, to foreigners in 2016, a 163% increase on 2015. By comparison, the United States issued around 1 million.

Yet Wang is optimistic. Last September, China's cabinet issued a notice saying it will be easier for students in China and senior foreign professionals to gain work and residency permits after laws are reformed. Internal barriers that make it difficult for foreigners to directly apply for grant funding in the Chinese system, including the requirement for Mandarin to be used in proposals, are also gradually being loosened, says Wang. ■

Hepeng Jia *is a science writer in Beijing and Ithaca, New York.*

IN THE FAST LANE

The past 15 years have seen China leapfrog Japan and the European Union in terms of total spending on research and development (R&D). China is expected to overtake the United States by 2020.



WHERE TO LOOK

A practical guide to finding a job in China

You can enquire after a job in China by applying for an advertised position, or contacting the university directly about available posts. If you have a target institution in mind, do some detective work to find out more about the university's hiring needs and who you need to impress.

Getting insider information and establishing connections is very important in China. Start by talking to the department chair in your discipline, then seek the advice of those already working in your chosen department, ideally those with a recent overseas background. Sometimes your peers will know more than senior faculty members about their team's hiring needs. Also try contacting former peers already working in China. If you're feeling daunted, remember that researchers in China are often encouraged by employers to help locate science talent and are sometimes offered financial rewards if they are successful.

If you're not sure where you want to work, check out the general job listings in science journals, advertised at conferences, and popular science websites. Don't forget Chinese websites too, such as ScienceNet.cn. Students should look out for opportunities to visit campuses in China. Some university recruiters will organize all-expenses-paid trips over holiday periods for prospective candidates to discuss jobs and opportunities.

Finally, bear in mind that not all Chinese universities are well equipped to deal with the red tape of recruiting a foreigner, such as dealing with work and residence permits and employment licenses. They may not even have English-speaking support staff. Foreign job seekers should approach institutions with a history of hiring foreigners if they want to be assured they'll receive the human resources support they need. **H.J.**

WHAT IS CHINA'S THOUSAND TALENTS PLAN?

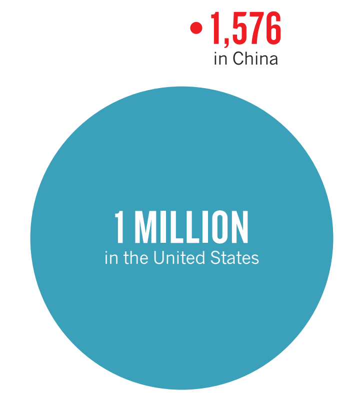
The nation's bid to lure back ex-pat scientists and recruit highly-skilled foreign researchers is now in its tenth year.

BY HEPENG JIA

In 2008, China's central government announced the Thousand Talents Plan: a scheme to bring leading Chinese scientists, academics and entrepreneurs living abroad back to China. In 2011, the scheme grew to encompass younger talent and foreign scientists, and a decade later, the Thousand Talents Plan has attracted more than 7,000 people overall. For Chinese scientists, the scheme has given them a strong financial incentive to return home. For foreigners, it's an opportunity to join the Chinese system with major administrative hurdles removed.

How does it work?

Previous programmes to lure back ex-pat scientists elevated postdocs and junior faculty members to full professorial or equivalent positions in China, but the Thousand Talents Plan was more ambitious, aiming to target professors and chief scientists in the West. This approach led to tougher recruitment criteria across all schemes, and high-level positions became less easy to come by. It's well understood in the research community that scientists recruited through talent schemes will gain access to much higher salaries and research funding levels than their locally trained peers.



Although China is keen to recruit foreign scientists, it remains difficult to become a permanent resident in the country. China issued **1,576** permanent residencies to foreigners in 2016, compared with approximately **1 million** in the United States.

Where do I begin?

To apply for any of the Thousand Talents schemes, you must already have a firm job offer from a Chinese institution. If you are a full professor at a leading Western university, you will more than likely be eligible for a place on the Thousand Talents scheme for senior academics, as long as your research record meets the demands of your field of study in China. The scheme is open to Chinese scientists under 55 years of age, and foreigners younger than 65. All applicants must have worked at renowned universities outside China, and have a strong publication record.

When you've confirmed a position with a Chinese institution, your university will probably suggest that you apply for the Thousand Talents programme, and ask for your CV with a list of published papers, copies of all academic qualifications (original or certified) and full-text copies of your highest impact research papers. Usually, letters of reference are not required. The university will then apply on your behalf. The length of the application process (which includes an interview) can be a matter of months if the applicant meets all of the criteria first time around. However, after signing their employment contract and agreeing to the terms of the Thousand Talents Plan, such as a minimum three to five years of working in China, scientists may need to wait for all aspects of their agreed research project to be realized, as not every financial and administrative decision is controlled directly by the university.

What are the benefits?

All successful applicants can expect a 1 million yuan (US$151,000) starting bonus, and the opportunity to apply for a research fund of 3–5 million yuan. Foreign scientists receive additional incentives, such as accommodation subsidies, meal allowances, relocation compensation, paid-for visits home and subsidized education costs. Employers are also obliged to find jobs for foreign spouses, or provide an equivalent local salary. In addition, the Thousand Youth Talents Plan targets foreign and ex-pat Chinese scientists under the age of 40 — unlike the main Thousand Talents Plan, Chinese returnees receive the same benefits as foreign recruits under the Thousand Youth Talents Plan. More details for all programmes can be found on the official website, 1000plan.org/en/.

"YOU MUST ALREADY HAVE A FIRM JOB OFFER."

Whether you were born in China or come from abroad, all applications to the Thousand Talents scheme go through your Chinese university employer. Only one institution can apply on your behalf, so you cannot make multiple applications.

What makes for a successful Thousand Talents application?

Your publication record, career record and communication skills are all important. Chinese institutions pay the greatest attention to the number of high-impact papers published. They also prefer scientists who have graduated from leading universities. Finally, make sure you establish the value you can add to your chosen research institution in China, highlighting your research ideas and particular areas of expertise.

A common mistake is to focus on one's own research without thinking about the department or school's research conditions and directions. Another is to stress the significance of your work, rather than how it could attract funding or when it could be published. In China, it is expected that newly employed scientists will be productive as quickly as possible. Creativity and originality are not valued as highly as publication rate. ■

Hepeng Jia *is a science writer in Beijing and Ithaca, New York.*

CHINA'S AI DREAMS

Why China's plan to become a world leader in the field of artificial intelligence just might work.

DANIELE MATTIOLI

BY OWEN CHURCHILL

Last year, China's chief governing body announced an ambitious scheme for the country to become a world leader in artificial intelligence (AI) technology by 2030. The Chinese State Council, chaired by Premier Li Keqiang, detailed a series of intended milestones in AI research and development in its 'New Generation Artificial Intelligence Development Plan', with the aim that Chinese AI will have applications in fields as varied as medicine, manufacturing and the military.

These AI ambitions, made public in July 2017, came as little surprise. 'Innovation' has been a favourite buzzword of China's leadership for several years, as the country seeks to transition from a production powerhouse to a centre of knowledge creation.

But China's AI aspirations are as economic as they are political, coming at a historic stall in the country's previously rapid growth — in 2016, the economy grew at its slowest rate since 1990. By 2020, the State Council's plan forecasts, the value of China's core AI industries should have exceeded 150 billion yuan (US$22.7 billion), and the total for all related industries should be 1 trillion yuan. By 2030, it is hoped those figures will be 1 trillion yuan and 10 trillion yuan, respectively.

Much suggests that China is already on the right track. In 2014, the country overtook the United States in terms of the number of research publications it produced — and, crucially, the number of those that were cited — on the subject of deep learning; in the past two years, Chinese teams have dominated the prestigious ImageNet AI contest, in which researchers compete to see which algorithms can best recognize images; and Beijing-based technology giant Baidu, which in 2017 announced it was launching a deep learning research laboratory in collaboration with the Chinese government, says it will have self-driving vehicles powered by AI technology on Chinese public roads by 2020.

But in its quest to become a leader in technology that does the job of people, China is uncharacteristically low on one thing: people. According to a recent LinkedIn report (see go.nature.com/2jvdcxe), there were more than 50,000 people in China's AI workforce in the first quarter of 2017. In the United States, a country with less than one-quarter of China's population, that number was above 850,000. India is home to an AI workforce of more than 150,000.

"For the time being, talent remains a major bottleneck in China's advances in AI," says Elsa Kania, a Washington DC-based analyst specializing in China's emerging technologies and defence innovation. She says that it is a lack of experience as well as a lack of people that is afflicting the country's AI sector. Indeed, LinkedIn found that 38.7% of those working in China's AI sector have more than 10 years' experience, compared with 71.5% in the United States. That, she says, "will continue to necessitate active efforts to recruit foreign talent from Silicon Valley and elsewhere". In Beijing's tech hub Zhongguancun, steps have already been taken. In 2016, the local government made it easier for foreigners to gain permanent residency status, and in 2017 it introduced an advice service to support entrepreneurial newcomers with everything from Chinese company registration and taxation to finance and intellectual property rights.

Indeed, strengthening talent is considered a matter of "utmost

IMAGINECHINA/REX/SHUTTERSTOCK

Students in Hong Kong play Chinese chess against an artificial intelligence system.

importance" in the government's AI development plan, which calls for "accelerating the introduction of first-rate global talent and young talent, to create a top talent base for China's AI". The plan makes specific reference to the Thousand Talents scheme, a government initiative that offers attractive financial packages to overseas scientists, both foreign and Chinese.

DOLLARS AND DATA

A number of factors will bolster efforts to make China an attractive place for top-tier AI talent, the first and foremost being the amount of public funding available to researchers. The value of this is particularly evident in today's world, says Kania, at a time when "the current US administration has not prioritized increased funding for research and development in AI".

Shu Chang, who works at the Shanghai-based AI start-up DeepBrain, says that "China's entire research environment might be better than that overseas." DeepBrain recently transitioned from producing intelligent hardware to focusing on cloud-based services. Referring to both AI and other technologies, Shu says that doctorate scholarships and research grants are easy to come by in China, especially when compared with Europe and the United States.

One of those benefiting from government capital is Sören Schwertfeger, a German robotics specialist who relocated to ShanghaiTech University to conduct research as an assistant professor. He says that he has received much more generous research funding from his university than he would have from institutions anywhere else in the world. Although specific purchases of research equipment must be signed off by the university, Schwertfeger says that his employer's input on the nature of his work is minimal. "I have my academic freedom and can do research on what I want."

"TALENT REMAINS A MAJOR BOTTLENECK IN CHINA'S ADVANCES."

ShanghaiTech University was set up in 2013 by municipal authorities and the government-run Chinese Academy of Sciences, which said it was responding to "a nationwide call to put innovation at the heart of China's development". In short, says Schwertfeger, "Shanghai is spending lots of money". His department comprises a mixture of foreign staff and Chinese staff with international experience.

Alongside abundant capital is another resource crucial to any AI engineer: data. The most populous country in the world — home to 730 million Web users — is a massive, accessible, treasure trove of workable data that are generally available to companies. Although online data protection laws were tightened in 2017 to protect citizens' personal information, there are still legal means for companies to obtain data. ▶

For example, Shu says that social media platform Weibo will provide developers with access to users' information, and such data are generally more available in China than abroad.

Schwertfeger says that his campus has around 3,000 cameras, a resource that could be tapped to drive facial-recognition research projects such as working out average class attendance. "Of course," he says, "not to spy on the students." Leaders in the AI field continue to cite ease of data access in China as a major selling point. As put by Wang Haifeng, vice president of Baidu, "data is the lifeblood of AI, and in China, we have excellent data pools".

"WE ENCOURAGE OUR EMPLOYEES TO EXPRESS THEIR OPINIONS."

It's an observation shared by others in the field. He Yong, an executive at DeepBrain, says that compared with citizens in Europe and the United States, the Chinese people have a much more relaxed attitude to personal privacy. He, whose company created the first Chinese-language mobile phone assistant — before Apple launched a Chinese version of Siri — says that AI companies in China "have an easier time both collecting data and putting it to use". DeepBrain's leadership is confident that such an environment makes China an attractive destination for global talent (see 'AI projects to watch') — although the company currently has no non-Chinese staff in its team of around 20. Nevertheless, the firm favours those with international experience, says Shu: "Technologies within AI, machine learning and deep learning all have their origins in the United States and Europe after all."

LINKS TO THE UNITED STATES

Taking a more proactive approach to enlisting foreign talent is Baidu, a company that has grown from a search engine into a multi-limbed tech giant increasingly focused on AI technology. The company's AI wing is actively seeking talent from abroad, where, says Wang, "the top echelon of AI talent is concentrated". As well as actively recruiting from US universities, Baidu operates two research and development centres in Silicon Valley, California, and opened another in Seattle, Washington, in October 2017. Such centres, says Wang, are crucial to not just harnessing international talent, but also ultimately channelling it towards China.

"Due to the United States' position as a hotbed of AI talent, our presence there is a main recruitment channel and a great access point," says Wang. The company's California-based AI research centre employs around 200 scientists — both Chinese and non-Chinese — contributing to the company's research in AI fields such as machine learning, big data, computer vision and natural language processing.

Sören Schwertfeger and Jiawei Hou work on a robot at ShanghaiTech.

DANIELE MATTIOLI

Baidu's presence in the United States has also led to the adoption of a work culture more closely associated with tech start-ups in the West than the often-hierarchic nature of Chinese companies, says Wang, adding that English is the working language for both Chinese and non-Chinese employees in their US centres. "We have a flat structure," he says, "and we encourage our employees to express their opinions freely, to enjoy being challenged." ■

Owen Churchill *is a journalist living in London who writes on tech, culture, and media in China.*

AI PROJECTS TO WATCH

DEEPBRAIN

- Artificial intelligence (AI) company developing speech-recognition software focused on the Chinese market.
- Small team of around 20 now focused on cloud-based services.
- Recently received 32 million yuan (US$4.8 million) in finance.

IFLYTEK

- Ranked sixth in *MIT Technology Review*'s '50 Smartest Companies 2017'.
- Developed voice-recognition software used by hundreds of millions of people in China.
- Valued at 84 billion yuan on the Shenzhen Stock Exchange.

CAMBRICON

- Government-backed deep-learning processor chip.
- Received 10 million yuan in funding in 2017.
- Aims to challenge Google's AlphaGo, a program that plays the Chinese board game Go, using a computer that's a fraction of the size.

BAIDU'S DEEP-LEARNING LABORATORY

- Collaboration between Chinese government and Baidu.
- Launched in 2017 with an undisclosed budget.
- Multiple universities involved, including Beihang, which is known for military research.

BAIDU SEATTLE RESEARCH CENTER

- Baidu's latest US research and development facility, following two in Silicon Valley.
- Facility to focus on AI and cloud computing.
- Seeks to tap into the Seattle region's computing talent.

TENCENT SEATTLE AI LAB

- First US-based AI lab for Asia's most valuable company.
- The lab will employ around 20 people to work on areas such as natural language understanding.
- Lab to be headed by former Microsoft scientist Yu Dong.

MOVING TO CHINA

Offering both extraordinary opportunities and intense cultural change, the decision to take your career out East can be complicated, even if China is your home country.

BY HEPENG JIA

PICTURES LTD./CORBIS VIA GETTY

China-born Xiaolai Zhou has temporarily given up the job hunt in his home country. The assistant professor at the Mayo Clinic in Jacksonville, Florida, discovered that finding his way into a professorship wasn't quite as easy as he'd hoped, and now he's having second thoughts. "Having been out of China for seven years, I missed my home and really wanted to return. But there are many personal factors involved in that decision," says Zhou.

Like many others, Zhou thinks that professional opportunities in his native land have become as good as, if not better, than in the West. However, owing to intense competition, he has been able to secure neither a position at a leading university nor an equivalent salary and funding package. Now, he's thinking twice about what the move would mean for him. "This move would mean I'd be nearer family and friends, but the United States remains the world leader in scientific research," he says.

Zhou is one of tens of thousands of overseas-trained Chinese academics who are now considering coming home. Until the early noughties, China suffered a dramatic net talent outflow — of the 700,200 Chinese studying abroad between 1978 and 2003, only 172,800 of them returned — but the situation has gradually reversed. Over the past few years, 70–80% of Chinese students have come home after finishing their studies.

As China's scientific recruitment scene becomes more competitive, returnees and overseas scientists need to be confident that they're making the right decision.

Family commitments and the challenges posed by cultural differences are some of the major driving forces behind a desire to return. Guojun Shi, a cell biologist at the University of Michigan in Ann Arbor, says it can be tough for Chinese academics to compete against their Western colleagues. "In the Chinese workplace, it's professional to be self-effacing, stress the insufficiency of your work and be deferential. But in America, to win research opportunities and resources, you must always stress your ability," he says.

THE CHOICE TO COME HOME

Weiwei Deng was a tenured associate professor at Virginia Tech in Blackburg before joining the Shenzhen-based Southern University of Science and Technology through the Thousand Talents Plan in early 2017. "My daughter is seven years old. If we do not return now, it will be hard for my daughter to adapt to the Chinese education system later on."

Others are eager to move closer to their relatives. Xiangfeng Jing chose to return to China after finishing his postdoctoral research at Cornell University in Ithaca, New York, because he wanted to be near his parents in the long term — despite concerns about missing career opportunities in the United States (see 'Like-for-like'). "If you decide to return, the earlier, the better. China will not always have so many vacant positions waiting for you," said Jing.

But swapping one culture for another can be tough. Physicist Ying Yan used to work at Lund University in Sweden, where academics with children over three months old can drop them off at publicly subsidized day-care centres. Now based at Soochow University in Suzhou, China, Yan finds the lack of support for early-year child care frustrating — in Chinese culture, family members are usually relied on to take care of their young children. Yan's husband is a physicist at Shanghai University, so the couple depend on Yan's mother-in-law to take care of their two-year-old son. "If we hired a nanny, that's one of our salaries gone," she says.

Returnee families are also frequently caught out by rules that prevent Chinese citizens from holding dual citizenship. For parents whose children were born abroad and hold two passports, trips out of China can involve elaborate planning to abide by visa regulations. A US passport, for example, is too valuable an asset to relinquish: the United States is the most sought after destination among Chinese citizens looking to emigrate, for reasons such as air quality, food safety and education. And some Chinese scientists, who have worked in the United States and obtained residency permits through their employers, worry that a move back to China will cause them to relinquish their residency status. Losing the freedom to work and live in the United States would be a harsh blow if the move to China does not go well and they choose to return.

Another widespread complaint among Chinese researchers is that some employers fail to keep their promises. Scientists are reluctant to speak openly on this subject, due to the potential for negative professional ramifications, but off-the-record examples include the failure of

Beijing affords great opportunities, but is blighted by air pollution.

administrators to push through projects, such as setting up a lab, if they encounter resistance from other staff. In a country where an individual's *guanxi* — social clout — is key to getting things done, few people risk their professional future by offending others (see 'Need-to-knows'). Scientists also say the intense competition in Chinese research can drive unscrupulous behaviour. Neuroscientist Bai Lu of Tsinghua University in Beijing notes that colleagues are often reluctant to share their ideas for fear they might be stolen or published by others.

THE CHINA EXPERIENCE

Foreign scientists working in China say that assimilating into local culture is a tough but rewarding task. Richard de Grijs, a Dutch astrophysicist who works at the Kavli Institute for Astronomy and Astrophysics at Peking University in Beijing, points out that nothing can be done without strong relationships on a practical and personal level.

"IT TAKES EFFORT TO BUILD UP A NETWORK AND BE SEEN AS A RESPECTED COLLEAGUE."

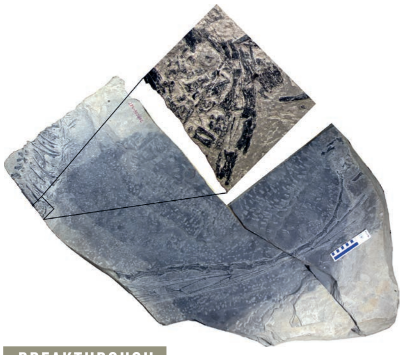
"It takes a lot of effort to build up a network and be seen as a respected colleague; that's not something you do overnight. I have established excellent relationships with a number of colleagues in the Chinese community, and we have become good friends on a personal level as well," he says.

Living in Beijing means that de Grijs is no stranger to air pollution. China's northern capital must often endure 'airpocalypses', or health-damaging levels of smog, which hit hard each winter. Despite government efforts to improve air quality by restricting factory emissions and toughening inspection procedures, China remains the world's deadliest country for outdoor air pollution, which kills more than one million people a year. The heavily industrialized and colder northern regions are the worst affected. "China's smog is a major problem and it's taking too long to improve," says de Grijs.

Scientists in the south are less affected by the dense smog. But there are other visible environmental concerns. For Joel Moser, a physics professor at Soochow University, the rising levels of water toxicity in Suzhou are worrying. Moser moved from Spain two years ago, where he was a researcher at the Institute of Photonic Sciences in Castelldefels. He says he was attracted by the opportunity of a professorship and higher levels of research funding. Suzhou is famed for its classically designed gardens and ancient water towns, but now the canals are heavily polluted. "Students tell me the dense algae in the water is still a recent trend. Some recall swimming in the nearby rivers not so long ago, when the water was still clear enough," he says. "Obviously, finding a balance between economic growth and environmental protection is challenging. But Chinese people are resourceful, and they will find a way."

A sense of anxiety about environmental living standards married with optimism for the future is common among foreign scientists. They point to the high levels of enthusiasm, talent and motivation of their students, warm reception from colleagues, and the affordability of smaller cities in comparison with their Western counterparts. "In the morning I often find fruits on my desk: a few bananas, an orange, and a handwritten note, which my students left as a token of appreciation," says Moser.

Naturally, the language barrier is a daily struggle, but it is possible to make it work, says Zach Smith, an American optics expert at the University of Science and Technology of China (USTC) in Hefei. Smith's decision to move was motivated by the offer of a 'dual-hire'. He and his Chinese wife Kaiqin Chu, who studies the same field, found it difficult to find jobs at ▶



BREAKTHROUGH

ANCIENT REPTILE GAVE BIRTH TO LIVE YOUNG

Pregnant fossil proves that not all archosaurs laid eggs.

BY JAMIE FULLERTON

Around 245 million years ago, just over 150 kilometres east of what is now the city of Kunming in southern China's Yunnan province, a pregnant *Dinocephalosaurus* died. In February 2017, the long-necked, flesh-eating marine animal and her fossilized embryo provided proof that members of her animal group could give birth to live young. Previously, researchers thought that the Archosauromorpha, ancient reptiles whose modern-day ancestors include crocodiles and birds, only laid eggs.

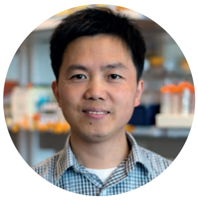
Palaeontologist Liu Jun and his team at the Hefei University of Technology in China published their analysis of the fossilized beast last year (J. Liu *et al. Nature Commun.* **8,** 14445; 2017). The specimen was one of 10,000 collected during 2008 in Luoping Biota National Geopark, an area that was long ago covered in shallow water. This is what helped to preserve the fossil after it died, says Liu.

Liu hopes that the finding will inspire other scientists to look for further evidence of live births in ancient reptiles, adding that studying fossils that have already been excavated could be as fruitful as searching for newly uncovered examples. He cites a 2011 *Science* paper showing live birth in a plesiosaur, a marine predator that went extinct around 66 million years ago. The plesiosaur specimen was excavated in 1987 in Kansas (F. R. O'Keefe & L. M. Chiappe *Science* **333,** 870–873; 2011). "We need to do more work with older specimens," Liu says.

Liu said that he received financial support from the China Geological Survey (CGS) and the National Natural Science Foundation of China. He suggests that although science funding levels in China are extremely healthy, they can be tied to the whims of leadership figures. China's former Premier, Wen Jiabao, used to be a geologist. When Wen retired in 2013, "geological research and investment went down, with less and less funding from the CGS", says Liu. ■

LIKE-FOR-LIKE

Two China-born former postdocs pursue careers on either side of the Pacific Ocean.



	MING LI ASSISTANT PROFESSOR UNIVERSITY OF MICHIGAN ANN ARBOR, USA	**XIANGFENG JING** ASSOCIATE PROFESSOR NORTHWEST A&F UNIVERSITY YANGLING, CHINA
NET MONTHLY SALARY	US$7,300	$1,600
RENT	$1,500 for a two-bedroom flat.	Employer provides free three-bedroom flat.
BENEFITS	Health insurance.	Health insurance. Wife offered job.
FUNDING LEVELS	$0.92 million start-up funding for three years, plus grants totalling $1 million (note that staff costs are much higher in the United States than in China).	$129,533 start-up funding for five years.
PROMOTION	Five-year tenure track.	Can apply for full professorship at any time.

the same institution in the United States, but USTC was more accommodating.

Describing his approach to grant applications, Smith says: "Typically I write the grant first in English, and then get a lot of support from students and Kaiqin to help translate it. But of course it's hard work because the language of grant applications is very specialized, and it's tough to keep the meaning while at the same time re-working the tone and flavour to suit the tastes of Chinese funding agencies."

"I WRITE THE GRANT FIRST IN ENGLISH, AND THEN GET SUPPORT TO HELP TRANSLATE IT."

Both returnees and foreigners learn to handle the role of politics in the life of a Chinese lab. Because most institutions are state-owned, all top-down decisions are understood to serve the needs of the ruling Communist Party. As such, it can be hard for scientists to openly confront perceived systemic failures without seeming to challenge the wider status quo. Scientists say they risk losing support for their work, and the chance for career advancement, if they speak out. They also complain that government bureaucrats make key decisions, rather than scientists, and expensive equipment purchases are made by unqualified administrators.

Foreign scientists report further troubles in China's system, including struggles with rigid Internet censorship practices. Websites including Google, Facebook and Twitter are blocked in China; to use key research tools such as Google Scholar and to access research papers, scientists use VPNs (virtual private networks) to climb China's Great Firewall. But VPNs are not always foolproof, as censors seek to block them.

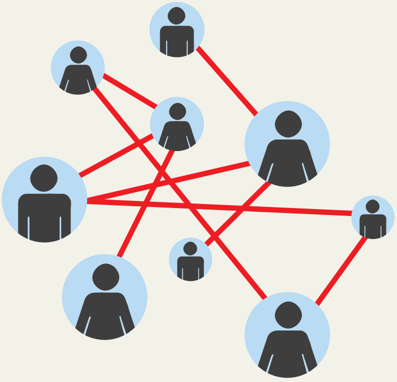
"This is partly why I did not choose to work in China full-time," says Olaf Wiest, who has been a visiting professor at Peking University for three months every year since 2010. "We must let potential job seekers have information on both sides, so that they can balance the gains and losses of coming to work in China." ■

Hepeng Jia *is a science writer in Beijing and Ithaca, New York.*

NEED-TO-KNOWS

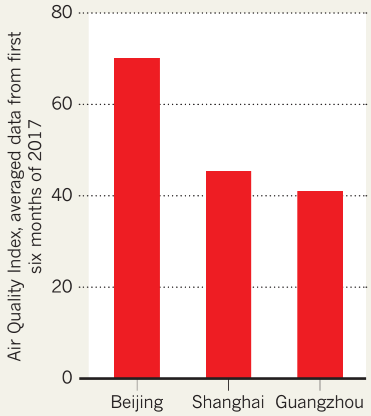
SOCIAL STANDING

Understanding *guanxi* — the personal relationships forged between individuals and their importance — is key for scientists looking to progress their careers in China.



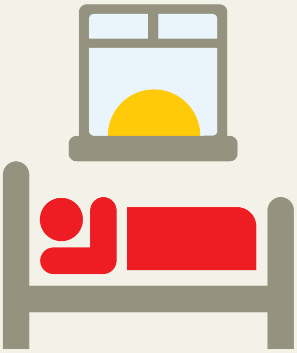
AIR QUALITY

Air pollution heads north and intensifies in the winter months, as coal use increases. But across the country, some cities enjoy much cleaner air than others.



TIME ZONE

China only has one time zone: Beijing standard time. This means that citizens of Urumqi, in the far west of the country, are often treated to a midnight sunset.



Beijing affords great opportunities, but is blighted by air pollution.

administrators to push through projects, such as setting up a lab, if they encounter resistance from other staff. In a country where an individual's *guanxi* — social clout — is key to getting things done, few people risk their professional future by offending others (see 'Need-to-knows'). Scientists also say the intense competition in Chinese research can drive unscrupulous behaviour. Neuroscientist Bai Lu of Tsinghua University in Beijing notes that colleagues are often reluctant to share their ideas for fear they might be stolen or published by others.

THE CHINA EXPERIENCE

Foreign scientists working in China say that assimilating into local culture is a tough but rewarding task. Richard de Grijs, a Dutch astrophysicist who works at the Kavli Institute for Astronomy and Astrophysics at Peking University in Beijing, points out that nothing can be done without strong relationships on a practical and personal level.

"It takes effort to build up a network and be seen as a respected colleague."

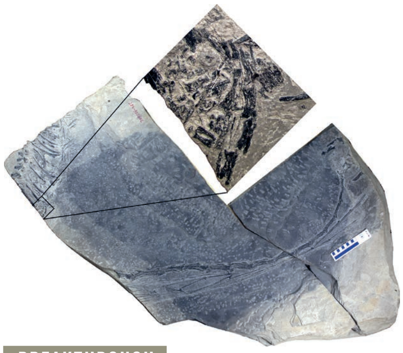
"It takes a lot of effort to build up a network and be seen as a respected colleague; that's not something you do overnight. I have established excellent relationships with a number of colleagues in the Chinese community, and we have become good friends on a personal level as well," he says.

Living in Beijing means that de Grijs is no stranger to air pollution. China's northern capital must often endure 'airpocalypses', or health-damaging levels of smog, which hit hard each winter. Despite government efforts to improve air quality by restricting factory emissions and toughening inspection procedures, China remains the world's deadliest country for outdoor air pollution, which kills more than one million people a year. The heavily industrialized and colder northern regions are the worst affected. "China's smog is a major problem and it's taking too long to improve," says de Grijs.

Scientists in the south are less affected by the dense smog. But there are other visible environmental concerns. For Joel Moser, a physics professor at Soochow University, the rising levels of water toxicity in Suzhou are worrying. Moser moved from Spain two years ago, where he was a researcher at the Institute of Photonic Sciences in Castelldefels. He says he was attracted by the opportunity of a professorship and higher levels of research funding. Suzhou is famed for its classically designed gardens and ancient water towns, but now the canals are heavily polluted. "Students tell me the dense algae in the water is still a recent trend. Some recall swimming in the nearby rivers not so long ago, when the water was still clear enough," he says. "Obviously, finding a balance between economic growth and environmental protection is challenging. But Chinese people are resourceful, and they will find a way."

A sense of anxiety about environmental living standards married with optimism for the future is common among foreign scientists. They point to the high levels of enthusiasm, talent and motivation of their students, warm reception from colleagues, and the affordability of smaller cities in comparison with their Western counterparts. "In the morning I often find fruits on my desk: a few bananas, an orange, and a handwritten note, which my students left as a token of appreciation," says Moser.

Naturally, the language barrier is a daily struggle, but it is possible to make it work, says Zach Smith, an American optics expert at the University of Science and Technology of China (USTC) in Hefei. Smith's decision to move was motivated by the offer of a 'dual-hire'. He and his Chinese wife Kaiqin Chu, who studies the same field, found it difficult to find jobs at ▶



BREAKTHROUGH

ANCIENT REPTILE GAVE BIRTH TO LIVE YOUNG

Pregnant fossil proves that not all archosaurs laid eggs.

BY JAMIE FULLERTON

Around 245 million years ago, just over 150 kilometres east of what is now the city of Kunming in southern China's Yunnan province, a pregnant *Dinocephalosaurus* died. In February 2017, the long-necked, flesh-eating marine animal and her fossilized embryo provided proof that members of her animal group could give birth to live young. Previously, researchers thought that the Archosauromorpha, ancient reptiles whose modern-day ancestors include crocodiles and birds, only laid eggs.

Palaeontologist Liu Jun and his team at the Hefei University of Technology in China published their analysis of the fossilized beast last year (J. Liu *et al. Nature Commun.* **8,** 14445; 2017). The specimen was one of 10,000 collected during 2008 in Luoping Biota National Geopark, an area that was long ago covered in shallow water. This is what helped to preserve the fossil after it died, says Liu.

Liu hopes that the finding will inspire other scientists to look for further evidence of live births in ancient reptiles, adding that studying fossils that have already been excavated could be as fruitful as searching for newly uncovered examples. He cites a 2011 *Science* paper showing live birth in a plesiosaur, a marine predator that went extinct around 66 million years ago. The plesiosaur specimen was excavated in 1987 in Kansas (F. R. O'Keefe & L. M. Chiappe *Science* **333,** 870–873; 2011). "We need to do more work with older specimens," Liu says.

Liu said that he received financial support from the China Geological Survey (CGS) and the National Natural Science Foundation of China. He suggests that although science funding levels in China are extremely healthy, they can be tied to the whims of leadership figures. China's former Premier, Wen Jiabao, used to be a geologist. When Wen retired in 2013, "geological research and investment went down, with less and less funding from the CGS", says Liu. ■

China's largest biobank at Zhangjiang High-Tech Park, Shanghai.

IMAGINECHINA/REX/SHUTTERSTOCK

BRACED FOR THE BIOTECH BOOM

Why careers in China's biopharmaceutical industries have never looked more promising.

BY SHANNON ELLIS

Timing can make a big difference in a career. Is it worthwhile to stay longer in a comfortable job or is it the right moment to strike out for a new challenge? Similarly, timing can make all the difference when deciding to enter a developing market like China.

Just a decade ago, when China-born scientists with overseas experience began returning to the country, lured by their homeland's fast growth and growing financial means, they found a drug industry dominated by generics. Undeterred, they got busy building the infrastructure for an industry capable of drug discovery and development, buoyed by substantial government support and a thriving economy.

Today, biotech specialists arriving in China find an industry at a turning point, with many key elements in place for innovation: a university system churning out doctorates and strong basic research, substantial financial backing from both the private and public sectors, regulations that are becoming globally harmonized and a vibrant group of entrepreneurial leaders with ambitions for China and abroad.

They also find a country facing significant unmet medical needs — particularly in cancer, neurology and diabetes — and a rapidly ageing population. Although China is the world's second largest pharmaceutical market after the United States, some of the most effective modern medicines are not on sale. For example, of the 42 cancer drugs approved globally in the past five years, only four are ▶

available in China. But this is set to change. Recent regulatory changes will bring imported drugs to China more quickly, and local biotechs are racing to develop domestic — and, they hope, global — blockbuster drugs. For academics and entrepreneurs, it is an ideal time to build on the biotech investments of the past, says Lan Huang, chief executive of New York-based BeyondSpring Pharmaceuticals, which is running drug trials in China.

A HUNGER FOR SCIENCE

It only took two visits to Shanghai's Zhangjiang Hi-Tech Park to convince Greg Scott to set up a life-science consulting business there, amid a hotbed of drug research and development (R&D) companies. He founded ChinaBio in 2007, and encourages others to consider a move to China. "Do it! It is a great experience," he says. "If I was helping someone plan their career, China has to be a part of it as the number one drug market outside the United States."

It is not just entrepreneurs and multinational drug company employees; many academic and staff scientists also find working in China a stimulating career move. Ray Stevens, a chemist renowned for determining the crystal structures of the body's receptors, which are important for identifying drug targets, can recall the exact moment he decided to trade sunny California for Shanghai, uprooting his school-age children and wife. Like many academics, he had visited China several times, but it was not until 2009, after delivering a talk on membrane proteins to colleagues in the neighbouring city of Suzhou, that he decided to make the move.

"One of the big attractions was the energy and excitement the students had for science. It won me over," Stevens says. After he had finished his talk, "a group of students came up to the podium to ask questions. They kept asking questions as I made my way to the bathroom and even followed me in. I was amazed; they were so hungry. It was the moment I decided to spend my sabbatical in China".

In 2011, Stevens moved to China as a visiting professor. Just a year later, the president of ShanghaiTech University, Mianheng Jiang, came calling, offering the chance to set up his own institute. He now runs the iHuman Institute at ShanghaiTech, is a member of China's Thousand Talents Plan and was in 2017 awarded a Magnolia Prize, an accolade given to foreigners who have contributed significantly to Shanghai's development. He has also co-founded a biotech company, RuiYi, in Shanghai.

BUILDING BIOTECH

The passion for science that Stevens discovered did not spring up accidentally. It has been fostered by government support for biotechnology that has intensified over the past decade, creating a force attracting scientists and the entrepreneurially inclined to China. Of the 2 million returnees to China over the past 6 years, it is estimated 250,000 work in the life sciences. And, although many scientists making the move were born and raised in China and have a decade or more experience working in the West, non-Chinese speakers such as Stevens and Scott are coming and thriving here, too.

> "SINCE 2008, 7,000 RETURNEES HAVE BEEN RECRUITED."

The push for innovation comes from the highest levels of government, with the biotech industry receiving special attention in not just one but three of the government's latest five-year plans: the strategic blueprints that determine the country's economic goals for the forthcoming half-decade. The latest plan, China's thirteenth, stipulates that the biotechnology sector should exceed 4% of gross domestic product by 2020 and that there should be 10 to 20 life-science parks for biomedicine with an output surpassing 10 billion yuan (US$1.5 billion). China has more than 100 life-science parks dotted across the country; run by local governments, these hubs lure companies with tax breaks and subsidies. It is estimated that more than $100 billion has already been invested in the life-sciences sector by state, provincial or local governments in an effort to hit the five-year-plan targets.

The Thousand Talents Plan (see page S8) has been especially successful at recruiting life-science talent. "Since 2008, 7,000 returnees have been recruited across all disciplines," says Dan Zhang, former secretary-general of the Thousand Talents programme and chief executive of Fountain Medical Development in Beijing, which helps companies to carry out clinical trials. "The life sciences committee for biotech is one of the largest groups in the programme. We've recruited more than 1,400 people, from both science and industry — including company founders, chief scientific officers or leading academics."

Returnees, especially those recruited via the Thousand Talents Plan, have had a "huge impact" on the industry, says Zhang. He says that returnees are the force behind the majority of drug approvals in China, that they fill peer review committees and life-science faculties, and that many are made university deans of schools of pharmacy and medicine. Sheng Ding, for example, has split his time between a biomedical-research facility in California and Tsinghua University in Beijing as dean of the school of pharmaceutical sciences since 2015. The generous grants and prestige of a place on the Thousand Talents Plan or similar programmes can increase an applicant's attractiveness to employers and enable them to command higher salaries.

DEEP POCKETS

Since the global financial crisis, financing for biotechs in China has been on the rise, whereas the sector has taken a hit in the West. Chinese investors who are looking to diversify their portfolios away from property and manufacturing are encouraged by the growth prospects of the life sciences, given China's unmet medical needs and ageing population. Chinese venture capital and private equity funds raised $45 billion for investment in the life sciences in the two and half years prior to June 2017, according to ChinaBio. So far, only $12 billion has been invested in the industry, with financiers on the hunt for good companies to invest in. Most of the cash is going towards financing innovative biotechs that

THREE YEARS

A snapshot of China's biotech industry

2015

2015, April

Chinese scientists alter a human embryo using CRISPR–Cas9 and trigger a global debate on gene editing.

2015, August

Regulatory reforms kick-off that will support the discovery and development of drugs.

2015, October

A Chinese biotech firm expands an existing deal with US pharmaceutical company Eli Lilly, adding US$1 billion in potential value to the partnership.

2016

2016, February–March

Two Chinese biotech firms successfully list on the NASDAQ stock market, raising more than $100 million.

FROM BIG PHARMA CAREERS TO RISKY START-UP OPPORTUNITIES
MEET CHINA'S BIOTECH STARS

FRANK JIANG, Asia Pacific R&D head and global vice-president at Sanofi until 2015

Joined CStone Pharmaceuticals in 2016.

LI CHEN, R&D head at Roche China until 2010

Founded Hua Medicine in 2011.

JINGSONG WANG, head of R&D at Sanofi China until 2015

Founded Harbour Biomedicine in 2016.

JIM WU, director at Roche in China until 2013

Founded Ark Biosciences in 2014.

SAMANTHA DU, a scientist at Pfizer until 2001

Founded Hutchison Medipharma in 2002 and Zai Labs in 2014.

STEVE YANG, head of R&D Asia for Pfizer, then Asia and Emerging Markets R&D head at AstraZeneca until 2014

Joined WuXi AppTec as executive vice president in 2014.

LINGSHI TAN, head of Pfizer China R&D and vice president of global development operations

Founded dMed in 2016.

are, in turn, hiring at a rapid pace (see 'Three years').

Money is flowing into academic life-science research as well. "It is relatively easy to get a good grant for science," says Xiaodong Wang, director of the National Institute of Biological Sciences in Beijing and co-founder of the immuno-oncology biotech BeiGene in Beijing. "Because the scale compared to the United States is still relatively small, in terms of relative money, and national young talents can get start-up funds from the central government, it makes things easier." But Wang points out that research grants are usually submitted in Chinese, which can be a barrier for overseas scientists, who have to rely on translators. Wang himself returned to China from the United States in 2003.

Stevens says he has not found accessing funding in China easier than in the United States, but is heartened by the emphasis on high-risk research versus the low-risk research that he says largely gets funded by the US National Institutes of Health. He credits Chinese grant-makers with investing for the long term, taking pressure off the need to generate data quickly to secure another round of funding. In China, academics are also increasingly able to profit from their research — universities are permitting inventors to share the proceeds from patents and set-up their own companies. This is part of a wider initiative to get more discoveries from the bench to the bedside.

BIOTECH BONANZA

Competition for life-science talent in China has shifted. Despite their heavy investment in R&D centres in China in the past decade, multinational biopharma firms are finding themselves battling Chinese biotechnology start-ups to attract talent. Two top pharma recruiters in Shanghai — Jonathan Zhu, head of life sciences in China for Heidrick & Struggles (also a returnee), and Simon Lance, managing director for China at Hays — predict that job growth will increasingly

2017

2016, July

A Chinese biotech start-up breaks the country's record for first-round financing, raising $150 million.

2017, June

China's drug administration body agrees to align Chinese drug regulations with the rest of the world.

Chinese CAR-T cell trial impresses on the global stage, showing excellent results for relapsed or refractory multiple myeloma.

ChinaBio reports that Chinese venture capital and private equity funds raised $45 billion over a period of 30 months for life-sciences investment.

Q&A: BIOTECH ENTREPRENEUR

Lan Huang founded BeyondSpring in 2010 and now splits her time between China and the United States.

BY SARAH O'MEARA

Why did you leave China?
I began my biology degree at Fudan University in Shanghai and transferred to the United States in my third year, in 1991. At that time, the government paid tuition fees and assigned you a job for five years after graduating. It was unlikely I would work in academia, and it was too long to be away from my research field, so I left.

Why did you move into biotechnology?
China joined the World Trade Organization in 2001, and I saw that as an opportunity. I borrowed money to start my own consulting firms in the United States and China. In 2010, I founded BeyondSpring, which is developing a drug to treat lung cancer and the effects of chemotherapy. We have offices in Dalian in northern China and New York. China has lots of high-quality data, which is a huge boost to our research.

How has China changed since you left?
It's incredible. Every time I visit, something has changed. When I left, there were no elevators, no central heating. Certain foods were rationed. Now, living standards have improved hugely. In many respects, it feels no different to America.

What's the difference between doing business in China and the United States?
Technically, they are increasingly similar. You can expect the same software, hardware, research methods and standards. The major difference is expectation. China is a developing country, and everyone is very commercially driven and practically minded. They want to get to the end fast. In the United States, it's more science driven. They enjoy the journey. ■

This interview has been edited for length and clarity.

▶ come from Chinese biotech companies, not the foreign biopharma firms. "It's a sea change," says Zhang.

The president and chief executive of Suzhou-based Innovent Biologics, Michael Yu, says that his company is expanding quickly. "We are constantly looking for employees and hiring people. We have grown more than 20% year-on-year." Innovent is hunting for employees who have worked in countries such as the United States, where the drug industries are more mature and people have had greater experience of overseeing the development of innovative drugs (see 'What recruiters want'). "Ten per cent of our team are from overseas," says Yu. "Returnees have first-hand experience with how drugs are developed and regulated in the United States." This type of foreign experience will become increasingly important. In July, China became a member of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), signalling its intentions to mould its regulatory system in the shape of the ICH's founding members: the United States, the European Union and Japan.

Start-up biotechs, especially those flush with cash from venture-capital financing, are looking to scoop up talent and are willing to pay top dollar. "In the last two to three years, we have seen a change. More R&D heads are considering offers to work for Chinese biotechs, venture capital and clinical research organizations," says Zhang. "I believe this will continue in the next few years as more investment flows into domestic start-ups and they can make a combined offering: money and equity." China-based biotechs such as BeiGene, Hutchison MediPharma, Zai Labs and WuXi Biologics have all enjoyed successful public listings, so other biotechs hope that equity offers will entice returnee talent away from corporate jobs and academic positions.

For many returnees, working for a start-up is not about the money; it is about increased responsibility and influence. Entrepreneurial biotechs demand more from their team than corporate environments, and returnees feel they can make an important contribution to building China's biotech industry. "In a large organization, one's role is minimal in terms of making an impact and being accountable for something that happens," says Xuefeng Yu, co-founder and chief executive of CanSino Biologics in Tianjin. "But in China, you can really contribute or lead in an effort that will be impactful for the industry and also society."

For Yu, deciding to leave an executive position with the pharmaceutical firm Sanofi in Toronto, Canada, to become an entrepreneur in China seemed like a reasonable risk to take. "I have plenty of experience and considered the chance to be successful to be pretty high," he says. Today, CanSino can point to its China Food and Drug Administration-approved Ebola vaccine, developed in partnership with the Chinese military, as a measure of its success.

Going directly from a corporate job to being your own boss in a new country is a big leap to make, but Yu was already deeply familiar with the Chinese vaccine market before he arrived, and was ready to do more. "I wanted to return to China to make products that would serve the country's health. It felt like the right time," he says.

A BIG ENTRANCE

For foreign scientists unsure whether China is for them, one well-trodden route is to first arrive in an expatriate role with a multinational pharmaceutical company, or to work with a recruitment firm offering positions overseas. Then, after a few years, many make the leap to a biotech start-up. The time in between provides an education on China. Returnees and foreigners may shine in technical roles, but there is a learning curve when it comes to understanding how to work with local staff and regulators, and run clinical trials. "I am always getting an education in managing talent in China," says Scott.

This path from multinational drug company to biotech is clearly seen in the bios of many of the more ambitious start-up founders (see 'Meet China's biotech stars'). Returnees with a few years under their belt in China and an ability to effectively navigate the system are far more valuable to employers than new arrivals.

CASE STUDY

What recruiters want

China lacks seasoned experts with at least 10 years' industry experience, say insiders. Companies are especially keen on experience in translational medicine, early-stage clinical trials and antibody manufacturing. Start-up biotech firms need experienced managers at every level, from clinical trials to drug-manufacturing processes, to help build their companies. Recruiters emphasize that those interested need to do their homework. "Come over, speak to companies and recruiters. Get involved in life science events and associations to get an indication about what China is and isn't," advises Simon Lance, managing director at the recruitment firm Hays. "You'd be surprised how blasé some candidates are when applying for positions here." S.E.

In less than a decade China doubled the provision of health insurance coverage from less than 50% of the population to 95%

"The advice I give when people first come back to China is: do not think you know China well just because you can speak Chinese," says Zhang. "People tend to underestimate the speed of change here and local capabilities. In the beginning, I tended to overemphasize my technical advantages and did not understand how to be fully effective in this environment." Or as Mark Engel, chief executive of Shanghai-based biotech firm Phagelux, puts it, he is looking to hire people with "skills plus a willingness to work within the culture".

UNIVERSITY NETWORKS

Academics looking to make a move to China can get their feet wet by attending conferences — and they may be able to exploit their existing networks of China-born students and lab mates. Stevens came to China at the invitation of his students, who sought help to set up their lab, and was offered a role as a visiting professor at the Shanghai Institute of Materia Medica during a sabbatical from Scripps Research Institute in La Jolla, California. He later went on to found the iHuman Institute.

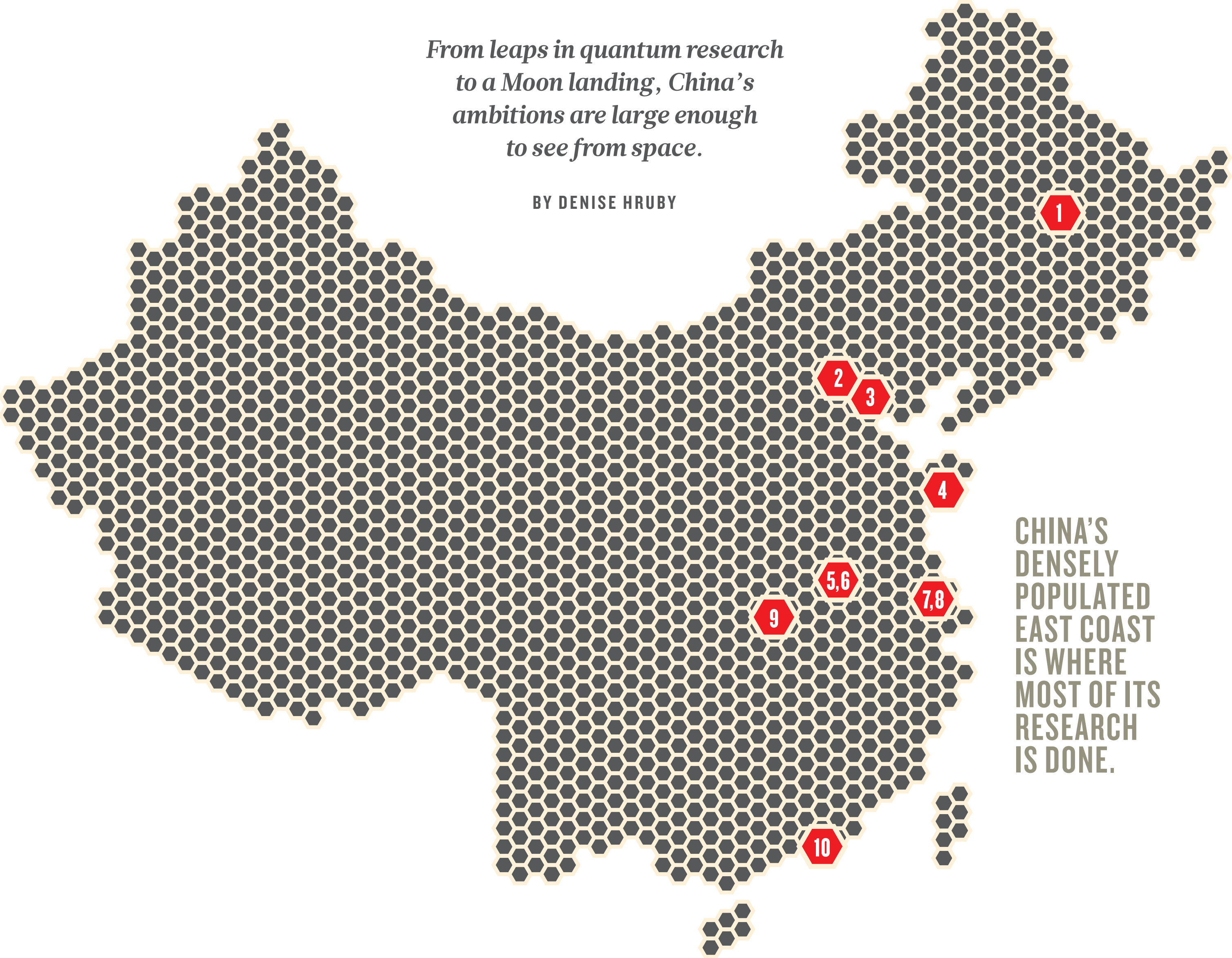
China's prestigious universities, such as Fudan, Tsinghua and ShanghaiTech, have had success in attracting international talent and want to open the doors further. Ming-Wei Wang, dean of Fudan University's school of pharmacy in Shanghai, wants the percentage of overseas faculty members to grow from 3% to 15% and to offer classes taught in English. Stevens has a similar goal for the iHuman Institute and has a target that 25% of his team is made up of overseas scientists. This push for global talent is driven by an understanding that strong science has no borders. But Stevens admits it is not always easy to get others to follow in his footsteps. "Getting non-Chinese foreigners to come to China has been a struggle. People are unaware of the opportunity and it takes an adventurous person to take the risk." ■

Shannon Ellis *is a science and business writer specializing in China and biotechnology.*

PUTTING SCIENCE ON THE MAP

From leaps in quantum research to a Moon landing, China's ambitions are large enough to see from space.

BY DENISE HRUBY



CHINA'S DENSELY POPULATED EAST COAST IS WHERE MOST OF ITS RESEARCH IS DONE.

When it comes to China's biggest achievements in science in recent years, many are quick to point to the world's first quantum-communication satellite. Launched in 2016, the spacecraft has already enabled researchers to prove quantum principles that could enable an unhackable communications network.

It's one of many visionary and large-scale research projects that China has launched with the aim of becoming a science and technology global leader by 2049, the centenary of the founding of the People's Republic of China.

According to research published in June last year, the satellite sent two entangled photons from space to ground stations more than 1,200 kilometres apart, as part of China's two-year Quantum Experiments at Space Scale (QUESS) mission, also called Micius after the ancient Chinese philosopher. The results demonstrated that the photons remained linked, providing proof that quantum entanglement can endure great distances. Entangled photons could theoretically serve as the basis for a completely secure communications network that hackers cannot penetrate without detection.

Scientists across the country were involved, from physicists working on the fundamentals at the University of Science and ▶

1 HARBIN INSTITUTE OF TECHNOLOGY

The Harbin Institute of Technology (HIT) is central to China's ambitious space-science programme, which includes the launch of the world's first quantum satellite, the establishment of a space station and a planned expedition to Mars. The team at HIT have worked on technology to propel craft into space using electrical rather than chemical energy. In 2016, China used this technology for the first time on an experimental satellite, Shijian-17.

2 NATIONAL SPACE SCIENCE CENTER

In 2016, Johann-Dietrich Wörner, director general of the European Space Agency in Paris, described China's National Space Science Center (NSSC) as dynamic, innovative and at the forefront of discovery. And Wu Ji, director general of the NSSC, was named one of China's top 10 scientists by *Nature* in 2016. Although China's work in space science began long after the United States or Europe, its achievements are impressive, including a 2013 Moon landing.

3 CANSINO BIOLOGICS

A private research laboratory based in the northern coastal city of Tianjin, CanSino Biologics was started in 2009 by a group of Chinese-Canadians who had spent their careers in high-ranking positions at major pharmaceutical companies before relocating to China. Currently, the company is working on a pneumococcal vaccine as well as common vaccines for meningitis and tuberculosis, and plans to have three vaccines on the market by 2019.

4 QINGDAO NATIONAL LABORATORY FOR MARINE SCIENCE AND TECHNOLOGY

China's first national-level marine science laboratory opened in 2015 in Qingdao. The marine sector in Qingdao — including trade, offshore oil exploration and equipment manufacturing — accounts for more than 20% of the city's gross domestic product. The lab has already made headlines with deep-sea explorations, and opened a joint research facility with scientists in Hobart, Australia.

5 DIVISION OF QUANTUM PHYSICS AND QUANTUM INFORMATION

Located at the University of Science and Technology of China in Hefei and established in 2003, the lab captured international headlines last year when its scientists successfully sent a pair of entangled photons from space to ground stations 1,200 kilometres apart, setting a new record. The team includes Lu Chaoyang, named as one of China's leading science stars in 2016 by *Nature*.

6 NATIONAL LAB FOR AI TECHNOLOGY

Opened in May 2017, the artificial intelligence (AI) lab at the University of Science and Technology of China in Hefei is the first national dedicated facility of its kind. At the time of its inauguration, Wu Feng, the lab's director, said his team would work on better understanding the brain's cognitive mechanisms, among other areas. The lab is also tasked with promoting the domestic development of emerging AI industries, such as robots. Several leading universities and Baidu, the Beijing-based Internet giant, are working closely with the team.

7 INSTITUTE OF NEUROSCIENCE

The Shanghai-based Institute of Neuroscience is at the heart of China's plans for brain science. In 2016, the government prioritized the discipline as part of its latest five-year plan. The focal point of research will be the China Brain Project, a multimillion dollar, 15-year project due to launch by the end of 2017. Scientists will explore how the brain works to develop treatments for neurological diseases and advance research into AI.

8 SHANGHAI SYNCHROTRON RADIATION FACILITY

When it opened in 2009, the synchrotron was the most expensive single scientific facility ever built in China: a total of 1.2 billion yuan (US$176 million at the time) was spent on the intense light generator, putting China into a small club of countries with similar facilities. In 2017, a collaboration between Chinese and Australian scientists at the synchrotron resulted in an algorithm to detect the formation of blood vessels, opening the possibility of detecting cancer much earlier.

9 WUHAN INSTITUTE OF VIROLOGY

The first lab in China to be cleared for high-risk work into the world's most contagious kinds of diseases is based at the Wuhan Institute of Virology in Hubei province, and will focus on researching Ebola. The Wuhan National Biosafety Level 4 Laboratory opened for trial runs in 2015, and in 2017 was approved to handle high-risk pathogens. By 2025, China aims to have at least five labs cleared to work on diseases caused by airborne organisms carrying infections.

10 BGI GENOMICS

This world-leading genomics research centre made its name after sequencing 1% of the human genome as part of an international collaboration. In 2007, the team moved to entrepreneurial tech hub Shenzhen, now home to Internet giant Tencent, tech giant Huawei and DJI, the world's largest drone maker. Analysts predict that BGI will be worth 20 billion yuan by 2020.

Q&A: RETURNEE EXPERIENCE

Materials chemist Yue Hu returned to Wuhan, China, to work at the Huazhong University of Science and Technology (HUST) after completing her PhD at the University of Edinburgh, UK.

BY SARAH O'MEARA

How did you find living in Edinburgh?
The city is small and beautiful, yet it is also a global place. I like the diversity of people who live there. I was surrounded by a lot of intelligent, and super friendly, people from many cultures. Interesting collaborations happened all the time.

Why did you decide to move back to China?
China has developed very fast in recent years and there are a lot of opportunities for young people, especially young researchers. I would love to visit other groups abroad in the future. It's always good to have more collaborators and friends, and to learn from other people.

What differences have you noticed?
People work longer hours in China. In Edinburgh, we were not allowed to go to the lab after 5 pm or during weekends. At HUST, it is very common for students to work from 9 am to 9 pm, and group meetings are regularly held on Saturdays and Sundays. Also, it's cheaper to do experiments in China: the cost of lab materials is generally lower, and there is also more funding available for researchers to buy equipment.

What do you like about living in Wuhan?
I like being close to my family and old friends. It is also in the centre of China, so it's convenient to go other places. Wuhan has many, many universities, which means there are lots of students and it feels very young and dynamic. ■

This interview has been edited for length and clarity.

▶ Technology of China in the growing science hub of Hefei to those receiving the photons at observatories in the northwest city of Qinghai and southerly city of Lijiang. The project was impressive not just because it could revolutionize the way humans interact, but because it was an entirely Chinese effort. "This research ranges from fundamental science to practical application, and it is a real successful model made in China," said Zhiyong Tang, professor of materials science at the National Center for Nanoscience and Technology in Beijing.

WORLD LEADING

When Micius's results were announced, international physicists recognized that China is becoming a world leader in quantum-satellite technology. A second satellite is already planned, as well as the largest ever quantum research facility, which will, among other research, work on stealth technology for submarines that will make them harder to detect.

Other space-science projects have also made headlines. The unmanned lunar spacecraft Chang'e-3 launched in 2013, and made China only the third country to land on the Moon. And 2016 saw the launch of the nation's most powerful rocket to date, the Long March-7, which will eventually supply China's planned space station. The country also launched its first X-ray telescope last year.

> **"ONCE YOU DEVELOP THAT REPUTATION, EVERYONE WANTS TO COME. SO THEN YOU GET THE BEST."**

In the remote southwestern province of Guizhou, China has finished constructing the world's largest single-dish radio telescope. The half-a-kilometre-wide dish, known as the Five-hundred-meter Aperture Spherical radio Telescope (FAST), was lauded as a unique feat of civil engineering. It is being used to detect radio messages from across the Universe to research phenomena such as dark matter and black holes, as well as listen out for indications of extraterrestrial life.

But China's scientists aren't just interested in what's above. *Jiaolong*, the country's first manned deep-sea submersible, reached a depth of 7,020 metres in 2012, a record for a scientific research vessel. In 2020, it will go on a one-year scientific research mission along the Pacific, Atlantic and Indian Ocean beds.

GRAND SCIENCE DESIGNS

There is every reason to think that China will continue to fund such grand projects for many years to come. Last October, President Xi Jinping told officials at the Communist Party's twice-a-decade national congress that the country must aim to reach new frontiers in science and technology. "We will strengthen basic research in applied sciences, launch major national science and technology projects, and prioritize innovation in key generic technologies, cutting-edge frontier technologies, modern engineering technologies, and disruptive technologies," he was reported as saying by Chinese state media.

In May 2017, the Shenzhen Grubbs Institute was launched at the Southern University of Science and Technology in Shenzhen. Backed by the Guangdong provincial government, the institute has launched a recruitment blitz to draw the "world's brightest minds". The institute, named after the 2005 Nobel Prize in Chemistry winner Robert H. Grubbs, will focus on the research and commercialization of drugs, materials and clean energy, beginning with plastics and polymers.

Grubbs praised the local infrastructure for rapidly commercializing discoveries in an interview with the state-owned broadcaster China Global Television Network. Reputation was key to attracting talent, he added. "You build a reputation of this department and this campus as really a great place to do science and research. Once you develop that reputation, everyone wants to come. So then you get the best." ■

Denise Hruby *is a writer and editor based in Shanghai, China.*

ONE APP TO RULE THEM ALL

In China, WeChat reigns supreme. The social messaging platform is used for everything from text chats and games to reading the news, paying bills and ordering taxis.

Visitors order from a 'restaurant of the future' at a conference in Hangzhou, China. The country has set aside around US$332.6 billion for innovative start-ups.

HUANG ZONGZHI/XINHUA/ALAMY LIVE NEWS

THE INNOVATION HUBS THAT DRIVE CHINA

China is transforming itself from a low-cost manufacturing base to a global industrial leader through the spread of commercial centres.

BY FLYNN MURPHY

A new 'innovation hub' seems to launch in China each week. Whether it's an office of one of the 'BAT' companies — Baidu, Alibaba and Tencent, China's most prominent tech giants — or a regional town's overnight conversion into a nerve centre for a particular hardware type or research field, few can deny the dizzying velocity of change and development. As of 2016, the nation had more than 1,600 business incubators.

"You're basically going from an agrarian society to a digital society without the bricks and mortar in the middle," says Hong Kong-based Warwick Pearmund, who works with Chinese tech companies as part of his role at the international recruitment firm Pure Search (see 'Q&A: Warwick Pearmund'). "That makes this region incredibly interesting." Pearmund advises that foreign researchers looking for work in China should think laterally about what the country needs next. For example, eyeing the nation's vast e-commerce system, now the world's largest, for opportunities in data science.

China's government is investing more money into its domestic start-ups than any other country in the world. In 2015, the total amount under management in government-backed venture funds was more than 2.2 trillion yuan (US$332.6 billion), according to data from Beijing's Zero2IPO Group. The money was held in around 780 so-called government guidance funds financed by tax revenue and state-backed loans.

In addition to making such huge investments, China has altered government regulations to help get projects moving, such as by relaxing visa requirements for staff and giving tax breaks to companies.

WHERE ARE THE COMPANIES?

Although innovation hubs are a dime a dozen in China, three cities are leading the pack in terms of commercially driven science and innovation: Beijing, Shanghai and Shenzhen. This is partly historical. Beijing has housed China's leadership almost continuously for six centuries and is home to its most prestigious universities, Peking and Tsinghua, which draw the most academically gifted students from across the nation each year. And it's the headquarters of leading Internet and search company Baidu, which is at the cutting edge of China's forays into artificial intelligence and machine learning. It is also a hub for Chinese returnees.

China's financial capital and most populous city, Shanghai, has the means and expertise to roll-out ambitious projects. A recent report by global consultancy firm KPMG on the world's top ten innovation hubs outside Silicon Valley and San Francisco ranked Shanghai as the number one to watch for the next four years, according to a survey of more than 800 technology company executives around the world. The authors credited the high-tech parks there and also noted that the city's "more pleasurable lifestyle and favourable climate" were attractive draws for outside talent. Although Shanghai's summers can reach temperatures of 40 °C, its air pollution is not as severe as Beijing's, which tied with Tokyo as the third city to watch, just after New York.

"Shanghai has been getting a lot of attention recently because the city's government is trying to make it the financial services hub of Asia," says Egidio Zarrella, a partner at KPMG China who focuses on innovation. He said the city would thirst for mathematicians and data scientists as it grows into this role.

China's third commercial centre is Shenzhen, which grew from a fishing village of around 30,000 people in the 1970s to become one of the world's biggest manufacturing hubs, with a population of more than 10 million. Shenzhen was China's first Special Economic Zone (launched in 1980 under the guidance of China's paramount leader Deng Xiaoping), growing from humble beginnings through government planning and foreign capital to become the world's factory. Now maturing beyond that status, diversifying into a centre for advanced manufacturing, robotics, genomics and more (see 'Q&A: Yongwei Zhang').

"Shenzhen is a very important hub for commercial applications," says Roy Green, outgoing dean of the Business School at the University of Technology in Sydney, Australia. "Every major corporate that wants to do very competent and relatively low cost prototyping comes to Shenzhen."

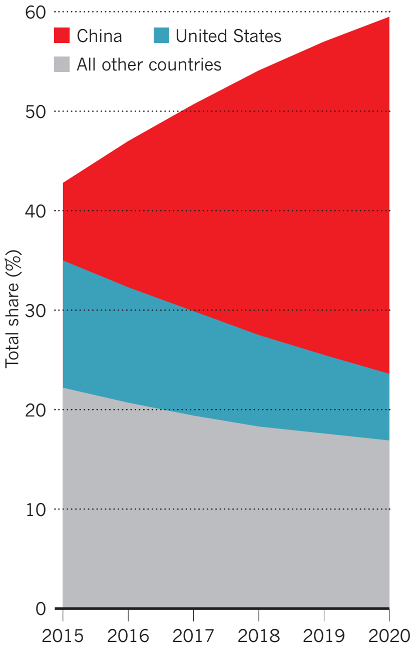
Mike Reed, an Australian mechatronics engineer who works at the high-profile start-up incubator HAX in Shenzhen, made the city his home almost three years ago. Reed says the efficiency of commercialization there is mind-blowing compared with its sluggish pace in his homeland.

"Living in Australia and making things don't really mix too well together," he says. The 24-year-old finds Chinese life affordable and convenient. He estimates that taxis cost one-quarter of the price in his native Brisbane. His monthly rent, nearly 4,000 yuan, is "pretty reasonable", and he uses apps on his phone to do everything from food shopping to paying his phone bill.

Reed is excited by the commercialization projects he sees every day. He cites a Canadian team of academics who are working on a highly efficient equipment-manufacturing tool called Pin Press, as well as graduates from the University of Pennsylvania in Philadelphia who are developing a high-pressure water jet, Wazer, that can cut through steel.

SALES SHARE

China's dominance of the world's e-commerce market is due to increase further over the next three years.



FROM INTENT TO INNOVATION

China's large financial investments in science and technology research and development (R&D), alongside ambitious targets for innovation, have begun to challenge the hegemony of the world's most technologically advanced nations. One-fifth of all money spent on all R&D comes from China, and the country plans to become a global leader in artificial intelligence by 2030 (see page S10).

Yet government targets are not always enough to kickstart innovation, says Green, citing Hangzhou, the home of Alibaba and the flourishing start-up ecosystem that sprang up around it. Now the world's largest e-commerce platform (see 'Sales share') and a shining example of Chinese innovation, Alibaba was launched from the apartment of former English teacher Jack Ma. "Ma picked up what was happening in Silicon Valley and developed this new operation from nothing. That has brought lots and lots of other start-ups," says Green.

Other cities hope to emulate Hangzhou's success, and the country's most powerful innovation hubs face stiff competition. In Shenzhen, Reed says that he has noticed the prices rising steadily in recent years, and David Zweig, a social scientist at the Hong Kong University of Science and Technology, says that "the cost of living is driving people out". Neighbouring, less-developed cities have been quick to seize an opportunity to attract talent. The nearby electronics manufacturing hub of Dongguan is working hard to increase its ability to ▶

Q&A: YONGWEI ZHANG

Executive director of BGI Research in Shenzhen, a non-profit human genome research organization, and chief operating officer of Complete Genomics in San Jose, California, a gene sequencing company owned by BGI Research.

What are the benefits of working in Shenzhen and at BGI?

People, money, infrastructure, ample flexibility, and a good salary package. 'Shenzhen speed' is much faster than in the United States and enables us to develop innovative products much faster.

And the drawbacks?

Chinese companies and employees work very hard, often at the cost of work–life balance. Also, the travel. At BGI, all employees, including senior executives, proudly travel in economy class.

How does Shenzhen compare to Silicon Valley?

Both Shenzhen and Silicon Valley are communities of immigrants and share the common goal of looking for better lives for themselves. This makes the cultures largely the same. But only recently has Shenzhen started to attract talent from other parts of the world. **FM**

This interview has been edited for length and clarity.

accommodate researchers, says Zweig.

China's central government also seeds the growth of potential commercial hubs. Xiongan New Area, a collection of towns and fields 100 kilometres southwest of Beijing, was last April designated as a new economic area. There has been talk of universities moving or expanding to Xiongan as it grows into an industrial centre: local media have reported early interest from Peking University. He Lifeng of the National Development and Reform Commission, which is overseeing the Xiongan project, said that science and tech innovation would be promoted in the city, and it has been touted as a commercialization hub that will complement the existing R&D infrastructure in nearby Beijing. Chinese state media have trumpeted its future significance with comparisons to the successful Special Economic Zones of Shenzhen and Pudong New Area in Shanghai.

Although the focus of China's development has long been the east and southeast coastal regions, the Belt and Road trade-route initiative, which will link China with Central Asia and Europe, promises to open up China's west, where cleaner air and a lower cost of living could entice businesses. Chengdu, the capital of the central Sichuan province, has been courting high-tech manufacturing: 300 Fortune 500 companies already operate there. And the impoverished southwestern province of Guizhou has launched efforts to turn its capital, Guiyang, and the surrounding area into China's 'Big Data Valley'. Tax incentives and government support have drawn Microsoft, Huawei, Hyundai Motor, Tencent, Qualcomm and Alibaba to set up offices in Guian New Area, a newly created urban and industrial zone an hour's drive from Guiyang, purpose-built to attract high-tech companies. The local government predicts that investment in the area will grow to $3.34 billion this year and add 30,000 jobs. Learning outsourcing company NIIT, based in Gurugram, India, announced in January that it would conduct training at the site, aiming to recruit and train 2,000 candidates per year. The courses will cover areas such as big data, cloud computing and cyber security, and will place candidates inside companies and government departments.

"If you have missed the investment opportunity in Guangdong or Zhejiang 30 years ago, by no means should you miss that of Guizhou today," Jack Ma was quoted as saying by the newspaper *China Daily*.

It's a pattern seen again and again in the world's most populous nation. There will be many more Guiyangs and Shenzhens as China continues its long-term transition to innovation-driven development, pouring capital and resources into history's largest and most ambitious industrial modernization project. ■

"THE WORLD DOESN'T REALLY KNOW JUST HOW ADVANCED CHINESE TECHNOLOGY IS."

Flynn Murphy *is a freelance health and science reporter based in Beijing.*

Q&A: WARWICK PEARMUND

Hong Kong-based associate director at Pure Search, an international recruitment firm.

What skills are Chinese companies looking for in outside talent?
It all comes down to data scientists. Whether mathematicians, statisticians or computer scientists. The rest of the world doesn't really know just how advanced Chinese technology is and how fast they can build businesses.

What gives potential recruits an edge?
Native Mandarin speakers who have either studied or worked overseas, particularly in Europe, the United States and Australia. In the last five years, it's noticeably changed — firms will say "bring us a native Mandarin speaker".

What do people dislike about working in China?
I have a client who's finding it frustrating working for a Chinese firm having worked overseas. You don't have the same freedom to make a difference. There's a lot more structure and it's more rigid. FM

This interview has been edited for length and clarity.

CAIA IMAGE/ALAMY

BREAKTHROUGH

SCIENTISTS REGENERATE LENS IN HUMAN EYE

Procedure removes damaged tissue to let stem cells grow.

BY JAMIE FULLERTON

In March 2016, it was revealed that a stem-cell therapy had given 12 Chinese infants suffering from cataracts the ability to see clearly (H. Lin *et al. Nature* **531,** 323–328; 2016). Lead scientist Kang Zhang, visiting professor at Sun Yat-sen University in Guangzhou and Sichuan University, said the regeneration of healthy lenses in children up to two years of age could be a paradigm shift in cataract surgery.

For five years, Chinese scientists worked in collaboration with researchers at the Shiley Eye Institute at the University of California in San Diego, where Zhang is a professor of ophthalmology, to develop a non-invasive surgical technique that can restore sight in just three months. During the procedure, surgeons remove the damaged lens from the patient but leave the lens epithelial stem cells intact. These grow and form a new lens to replace the old one.

Zhang says that conducting the clinical trials and tests on primates in China, rather than the United States, helped to keep costs low. He adds that attitudes towards animal testing in China, where animal rights protests are far rarer than in the United States, helped to move the research along quickly.

"If we were just doing it by ourselves in the United States it would have taken five to ten years," he says. "We were able to accomplish it in two to three years."

Praise came from around the world. Dusko Ilic, a stem-cell scientist at King's College London, called the research "one of the finest achievements in the field of regenerative medicine", and "science at its best".

Zhang says that further research using his team's techniques could further "harness a patient's own ability to regrow organs" in other areas of the body, such as the liver and brain. ■

DANIELE MATTIOLI

DRAWN TO SHANGHAI'S SCIENCE SCENE

Physics professor Cosimo Bambi was educated to PhD level in his native Italy and has held visiting positions at universities in the United States and Japan. But he says he really got into his research groove at Shanghai's Fudan University.

BY JAMIE FULLERTON

Although Fudan University — one of China's most prestigious — has its campuses in the heart of the country's most populous city, Cosimo Bambi says that Yangpu District in Shanghai feels more like a small town. "You can live there and not worry about many other things — several students live inside the campus, they don't have to leave," he adds.

Since moving to China, the 36-year-old professor has completed 63 research papers, covering gravity, cosmology and high-energy astrophysics. He has been at Fudan since 2012, having won a place there through China's Thousand Young Talents Plan, which is part of the wider Thousand Talents recruitment scheme launched by the Chinese government in 2008 to encourage highly skilled people — especially scientists — to work in the country.

Bambi has already stayed far longer than his programme's three-year minimum requirement, and his contract is now considered for renewal every three years; a process he describes as a formality. He says the Chinese government's keenness to recruit foreign talent has helped to create a working environment in which he feels he can achieve far more than he would in Europe or the United States. "I might be just a researcher working with one postdoc or young research student," he says of the United States. "Here I have several students and my teaching load is low: only one course per year. In Europe and North America you have to teach a few more classes."

He decided to move after speaking to Chinese colleagues at the University of Tokyo, who lauded their country's heavy focus on research. He hasn't looked back since. "It's completely different," he says. "For a foreigner, it's impossible to find a permanent position in Japan due to the country's relatively closed scientific culture. But in China, I don't feel like there is any difference in terms of opportunities."

Bambi's research team usually comprises around one to four postdoctoral researchers, as well as four or five PhD students and a large number of undergraduates. He says that the team works well, and that the challenges of managing it are very different to those he would face working in the United States or Europe.

"Typically Chinese students are hard-working, but having a good academic record is often completely uncorrelated with their research skills," he says. He adds that the Chinese education system can foster students who work very hard, but do not always show great initiative in their work.

Students are also required to take additional classes outside their core subject, he says. "A student of mine had to attend ceramics, sports and biology classes. All very interesting subjects, but probably better to be attended at high school, and not useful to physics."

Yet Bambi adds that his students are, on the whole, highly motivated in comparison with those he met in Europe. "In Italy, students are more relaxed. The admission process feels less competitive. Students tend to go to the university in their home town. And there's an expectation people are also there to have fun. Whereas at Fudan, students have gained top marks in a gruelling university entrance exam and, often, the weight of their family's expectations is upon them."

Language is a problem. He started learning Mandarin after moving to China, but gave up after a few months, realizing that to speak the language with any degree of competency would take a huge amount of work. Instead, while at work, Bambi is helped by his secretary. The rest of the time he relies on translation apps such as Google Translate. He pays his bills with the Chinese mobile app Alipay, and uses the messaging app WeChat to stay in touch with friends. Both have English versions.

"I lead a very simple life," he admits. "I live very close to campus, and eat in the canteen. My friends and family don't tend to visit as it's too far for them."

Bambi has travelled extensively in China, often combining a work trip with sightseeing. He particularly likes the southern cities that are surrounded by lush mountains, which offer a stark contrast to the flat, concrete cityscape of Shanghai. "The coastal city of Shenzhen, for example, is very lush, and surrounded by forested hills," he says.

At present, Bambi has no plans to leave China. His final decision, he says, will come down to balancing professional opportunities against concerns about poorer air quality and a far less generous welfare system than in Europe.

"For my salary, for my research, for my work, it is a good place," he says. "But there are many things to take into account," he says. "Who knows? Maybe I'll stay forever." ■

Jamie Fullerton *is a freelance journalist based in Asia. Additional reporting by* **Sarah O'Meara**.